# OVERVIEW OF APPROACHES FOR MISSING DATA

Susan Buchman
36-726
Spring 2018

# WHICH OF THESE PRODUCE MISSING DATA?

- A patient in a trial for a new drug dies before the study is over
- A patient in a trial for a new drug doesn't die for the next 30 years
- A car rental company near LaGuardia Airport in NYC requires your 5-digit zip code
- I don't answer a survey question that I was never shown because of "skip logic"
- I skip a survey question because none of the answers are relevant to me
- I'm asked to rate something on a scale of 0 to 100, and I answer -1
- Unintentional overwriting of 10 rows in a database table with NULL

# WHAT IS MISSING DATA?

- Some definitions are based on **representation**: Missing data is the lack of a recorded answer for a particular field

- Other definitions are based on **context**: Missing data is lack of a recorded answer where we "expected" to find one

As we'll see, lack of a recorded answer is neither necessary nor sufficient for being missing data.

# WHY DO WE CARE ABOUT MISSINGNESS?

Missing data can result in:

- Reduced statistical power
- Biased estimators
- Reduced representativeness of the sample
- Generally incorrect inference and conclusions

# THERE IS NO ONE CLEAR ANSWER FOR HANDLING MISSINGNESS

"All I know is that you throw out missing data and make a note of it."

"All I know is that you throw out missing data or fill it in and make an informative note of it."

# HIGH LEVEL AGENDA

- EDA for missingness
- Mechanisms for missingness
- Handling missingness
- Special cases

Our goal for today is to develop the **vocabulary** of missingness

# EDA FOR MISSINGNESS

- Quantifying and Visualizing Missingness
- Disguised Missingness

# DO YOU HAVE MISSINGNESS?

Treatments generally launch into discussion of *modelings* missingness. Need to first understand its scope.

Same principles as the rest of statistics:

- Easier to visualize if you have fewer variables
- Want to understand univariate and multivariate relationships

# HOW IS MISSINGNESS REPRESENTED IN YOUR DATASET?

- Don't assume it will be in native form
  - Blanks
  - Empty stings
  - NA
  - NULL
- Anything else that well-intentioned humans may come up with
  - -999999
  - "Did not answer"
  - "Ugh, sensor was broken"

"Disguised Missingness"

# BE AWARE OF DISGUISED MISSINGNESS

"When a standard code for missing data is either unavailable or its use will cause real or perceived difficulties for data entry personnel (e.g., angry words from a supervisor), data values are likely to be entered which are formally valid (i.e., exhibit the correct data type, satisfy edit limits, etc.) but factually incorrect..."

Ronald K. Pearson. 2006. The problem of disguised missing data. SIGKDD Explor. Newsl. 8, 1 (June 2006), 83-92.

# HOW IS MISSINGNESS REPRESENTED IN YOUR DATASET?

- Mixing of missing indicators, e.g. both NA and NULL in the same variable, may indicate different interpretations; don't necessarily collapse

- Don't hesitate to reach out to the client or other subject matter experts

# NULL DATA CAN BE VISUALIZED

**`naniar`** and **`Amelia`** in R can produce "missingness maps"

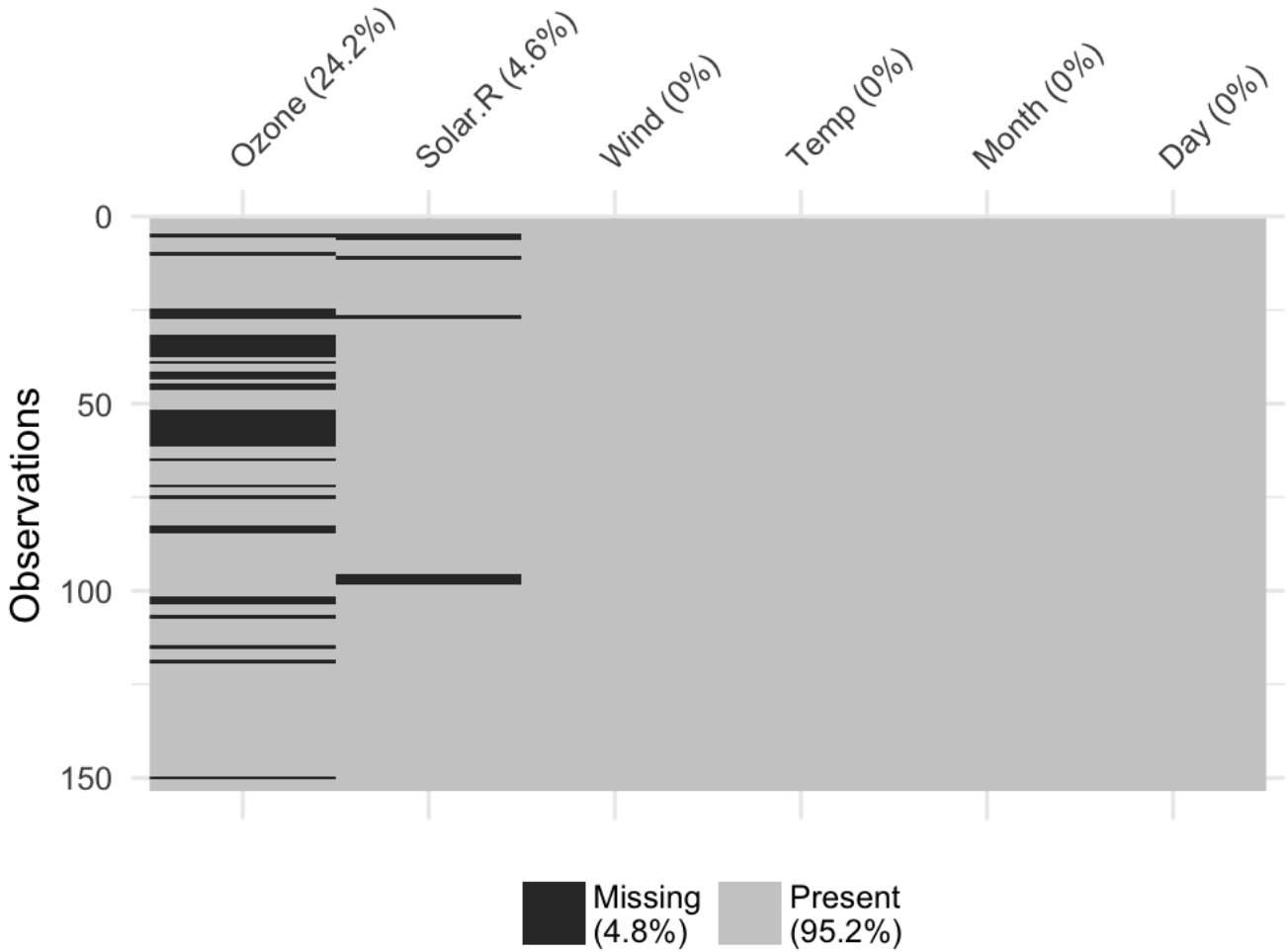
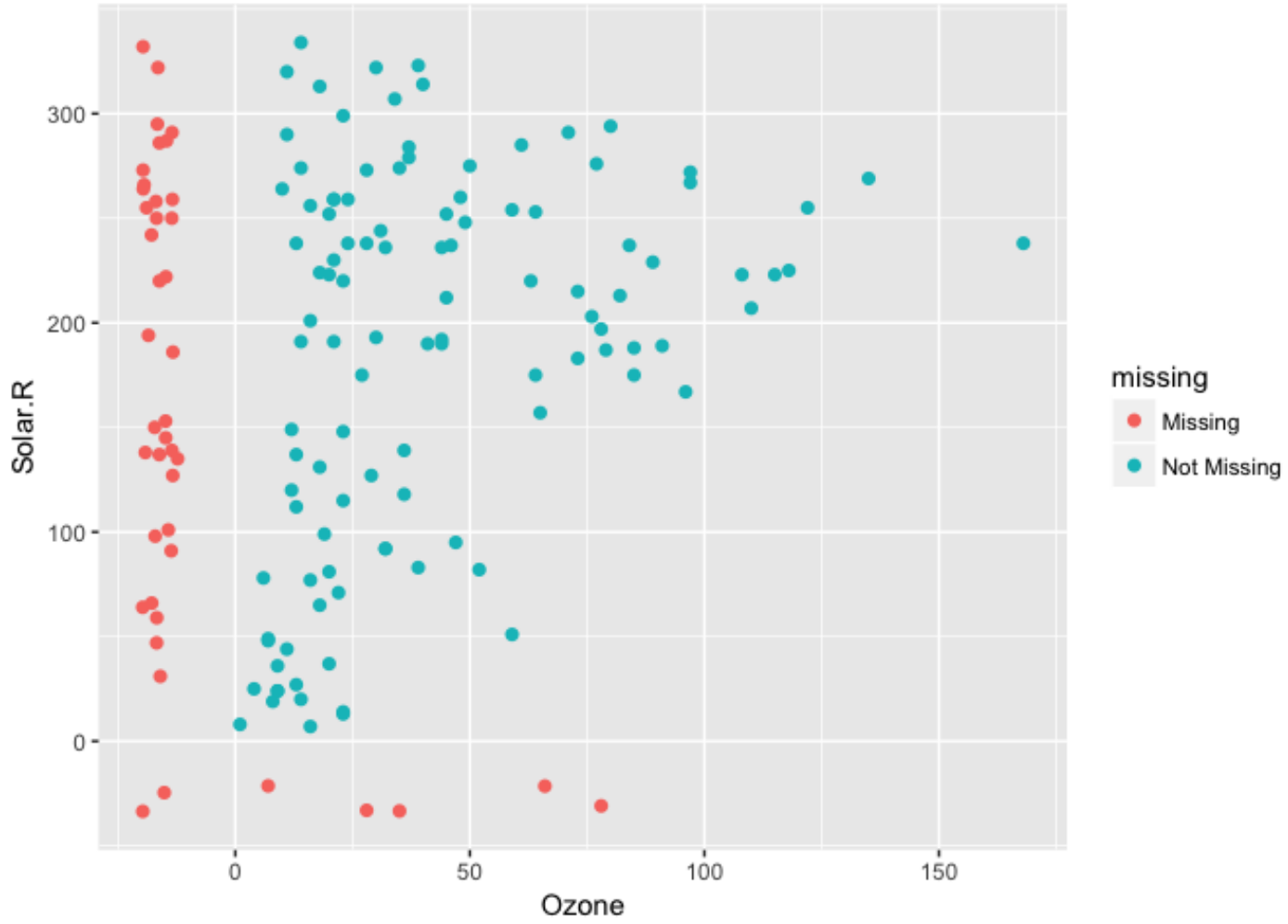



Image taken from http://naniar.njtierney.com/

# NULL DATA CAN BE VISUALIZED

**`naniar`** can produce missingness-grouped plots

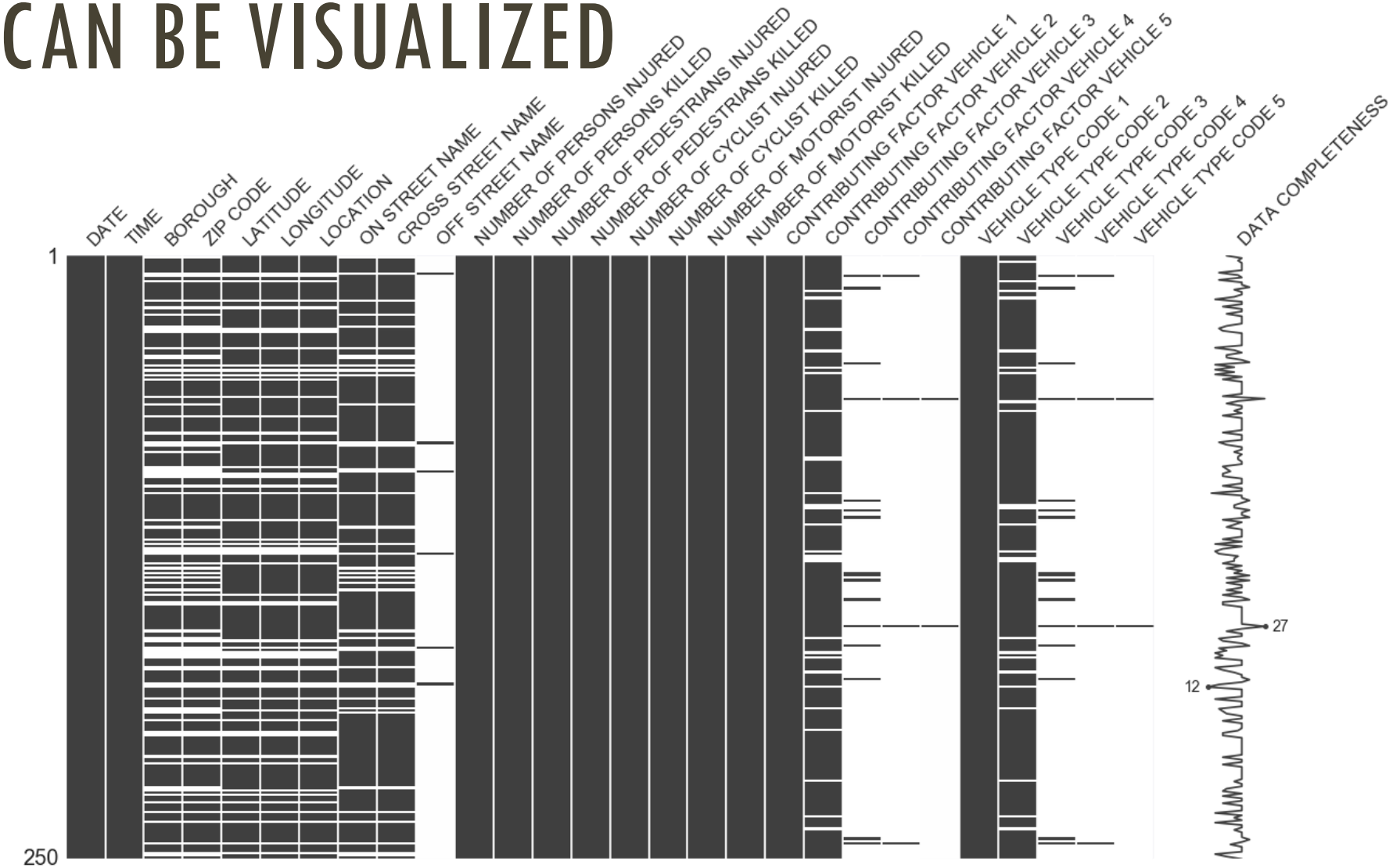

Image taken from http://naniar.njtierney.com/



# NULL DATA CAN BE VISUALIZED

**`missingno`**

in Python

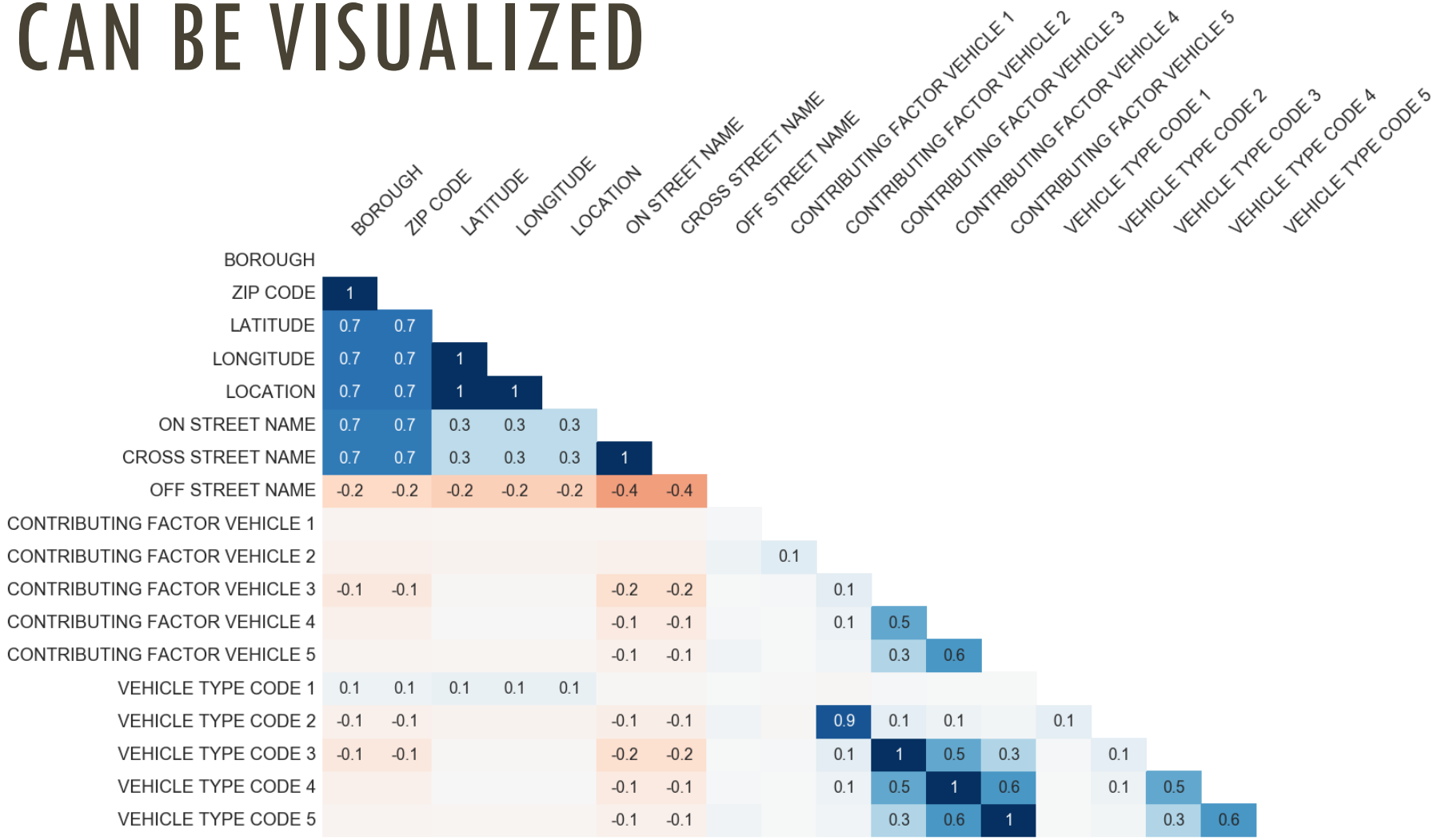



http://www.residentmar.io/2016/03/28/missingno.html

# NULL DATA CAN BE VISUALIZED

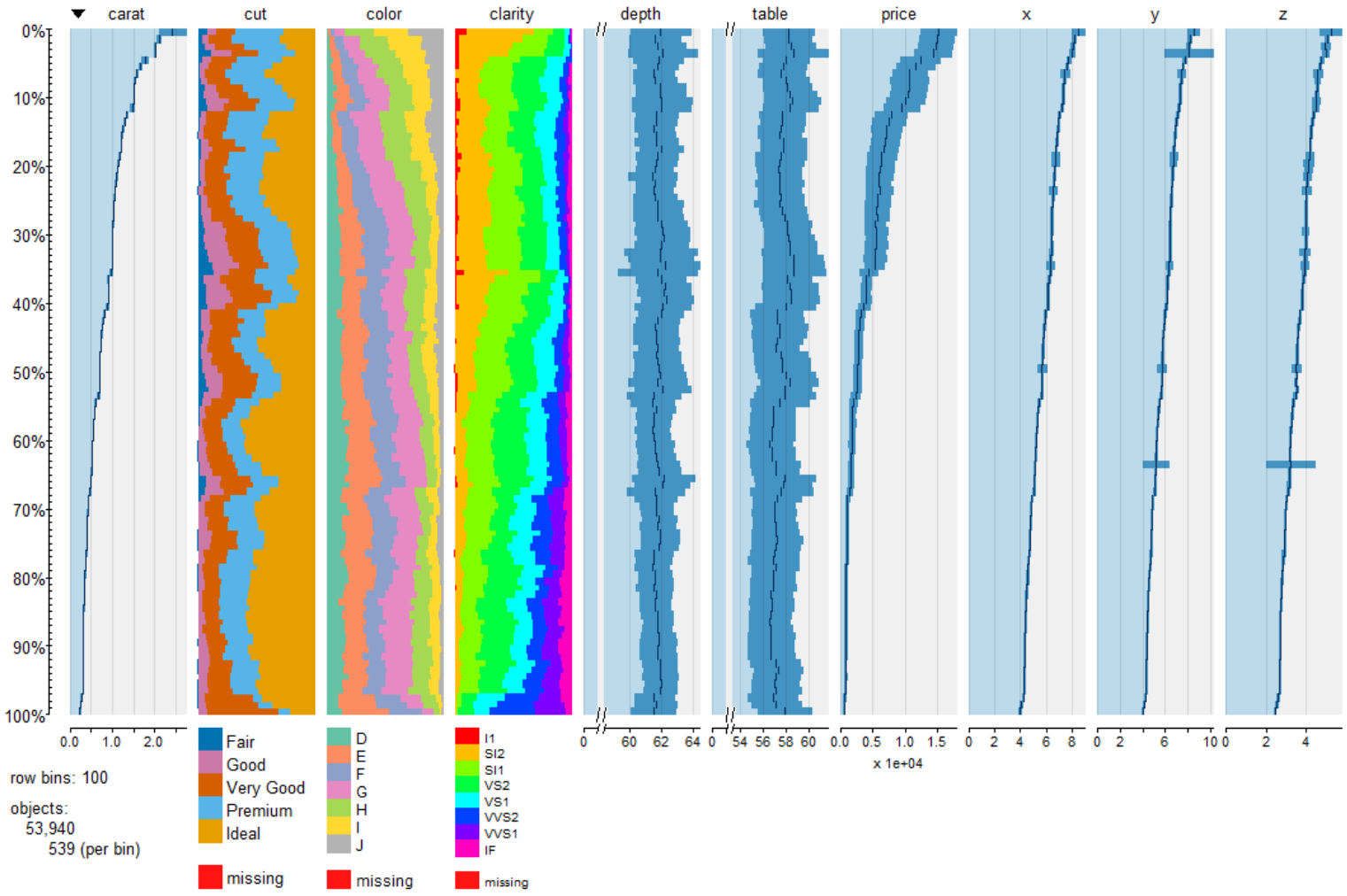
**`missingno`**

in Python

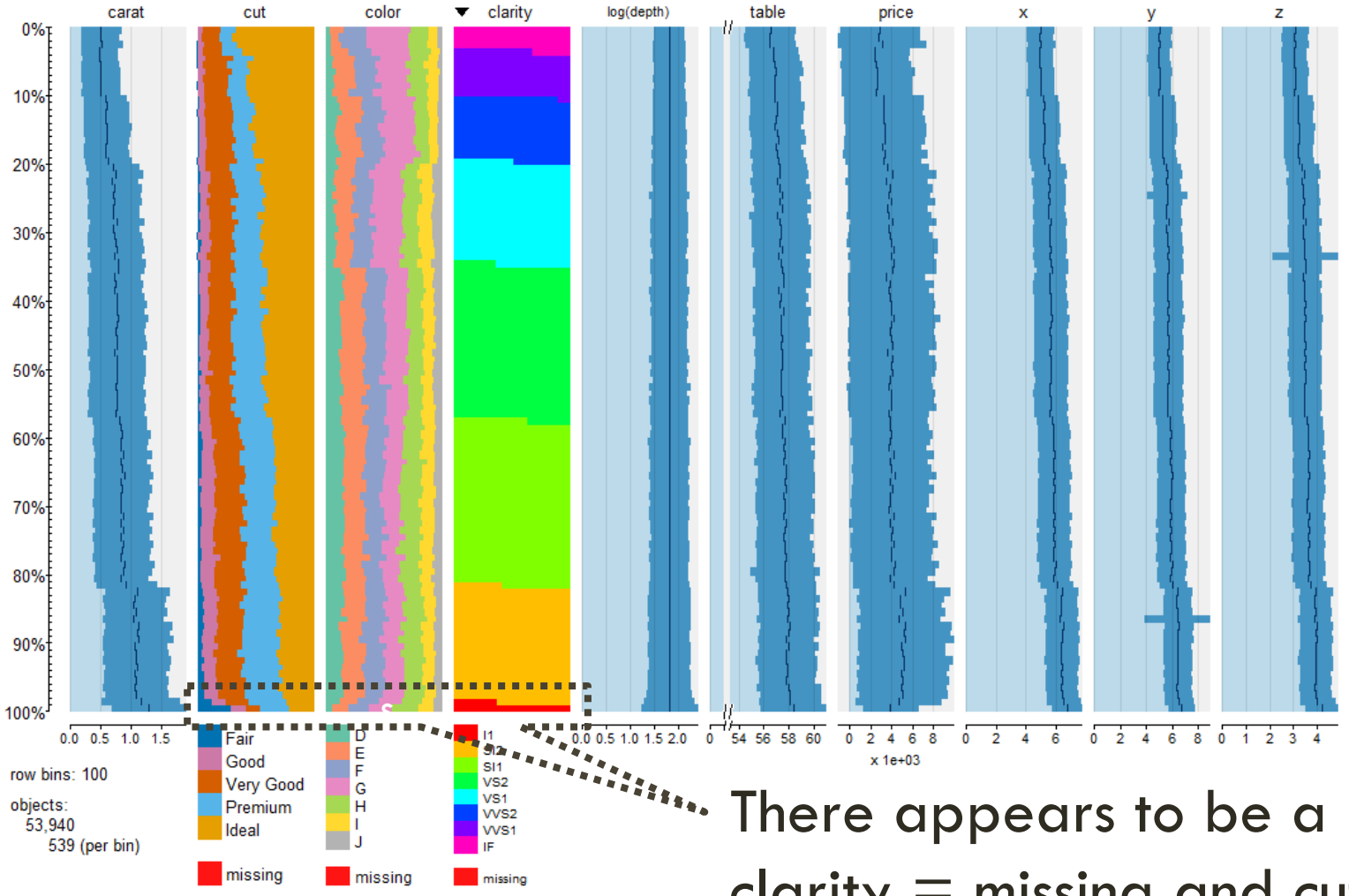



http://www.residentmar.io/2016/03/28/missingno.html

# EVEN GOOD OLD TABLEPLOT

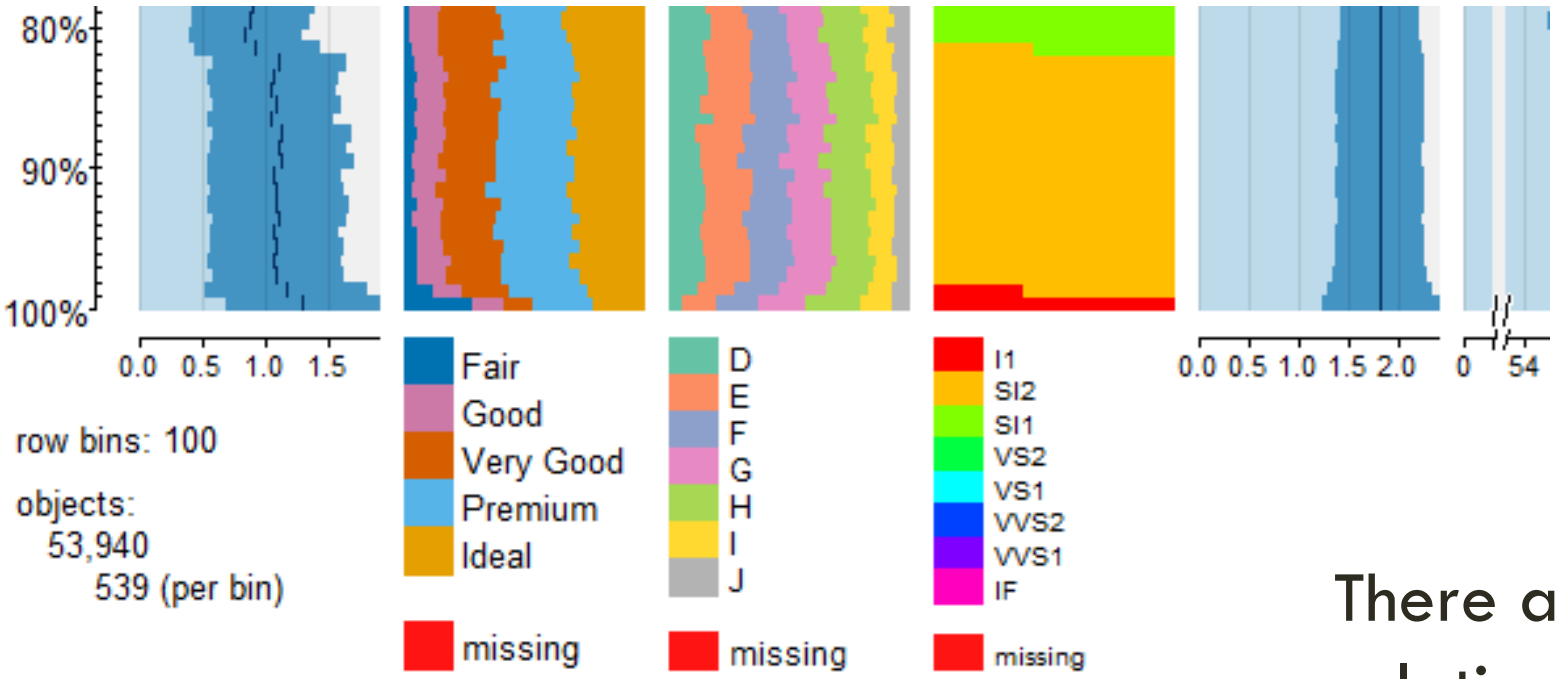


# EVEN GOOD OLD TABLEPLOT




There appears to be a relationship between clarity = missing and cut = fair

# EVEN GOOD OLD TABLEPLOT




There appears to be a relationship between clarity = missing and cut = fair

# MECHANISMS OF MISSINGNESS

# ARTICULATE WHY THE DATA ARE MISSING

1. **Missing Completely at Random (MCAR).** The data are equally likely to be missing.

2. **Missing at Random (MAR).** The likelihood of being missing depends only on non-missing data.

3. **Missing Not at Random (MNAR).** Missingness depends on unobserved data or the value of the missing data itself.

# THE ICE CREAM STUDY

I want to understand the ice cream preferences of MSP students.

I have record[illegible] gender of each student, and then ask what your favorite ice cream is. Consider the following missingness scenarios:

1. I ask all the students their preferences, but unbeknownst to me, people who prefer vanilla are embarrassed because it's so plain and refuse to answer
2. I ask all the students their preferences at random, but am interrupted after 30 students
3. I ask all the women their preferences at random, but am interrupted halfway through asking the men at random

# WE CAN FORMALIZE THESE DEFINITIONS

Let **X** represent a matrix of the data we "expect" to have; $\mathbf{X} = \{\mathbf{X_o}, \mathbf{X_m}\}$ where $\mathbf{X_o}$ is the observed data and $\mathbf{X_m}$ the missing data.

Let's define **R** as a matrix with the same dimensions as **X** where $\mathbf{R_{i,j}} = 1$ if the datum is missing, and 0 otherwise.

1. MCAR:
   $\mathbf{P(R \mid X_o, X_m) = P(R)}$
2. MAR:
   $\mathbf{P(R \mid X_o, X_m) = P(R \mid X_o)}$
3. MNAR:
   No simplification.

# IT IS OFTEN IMPOSSIBLE TO BE CERTAIN WHICH MECHANISM APPLIES

1. The good news is that we can test for MCAR!

   Roderick J. A. Little (1988)

2. The bad news is that we cannot test for MAR versus MNAR.

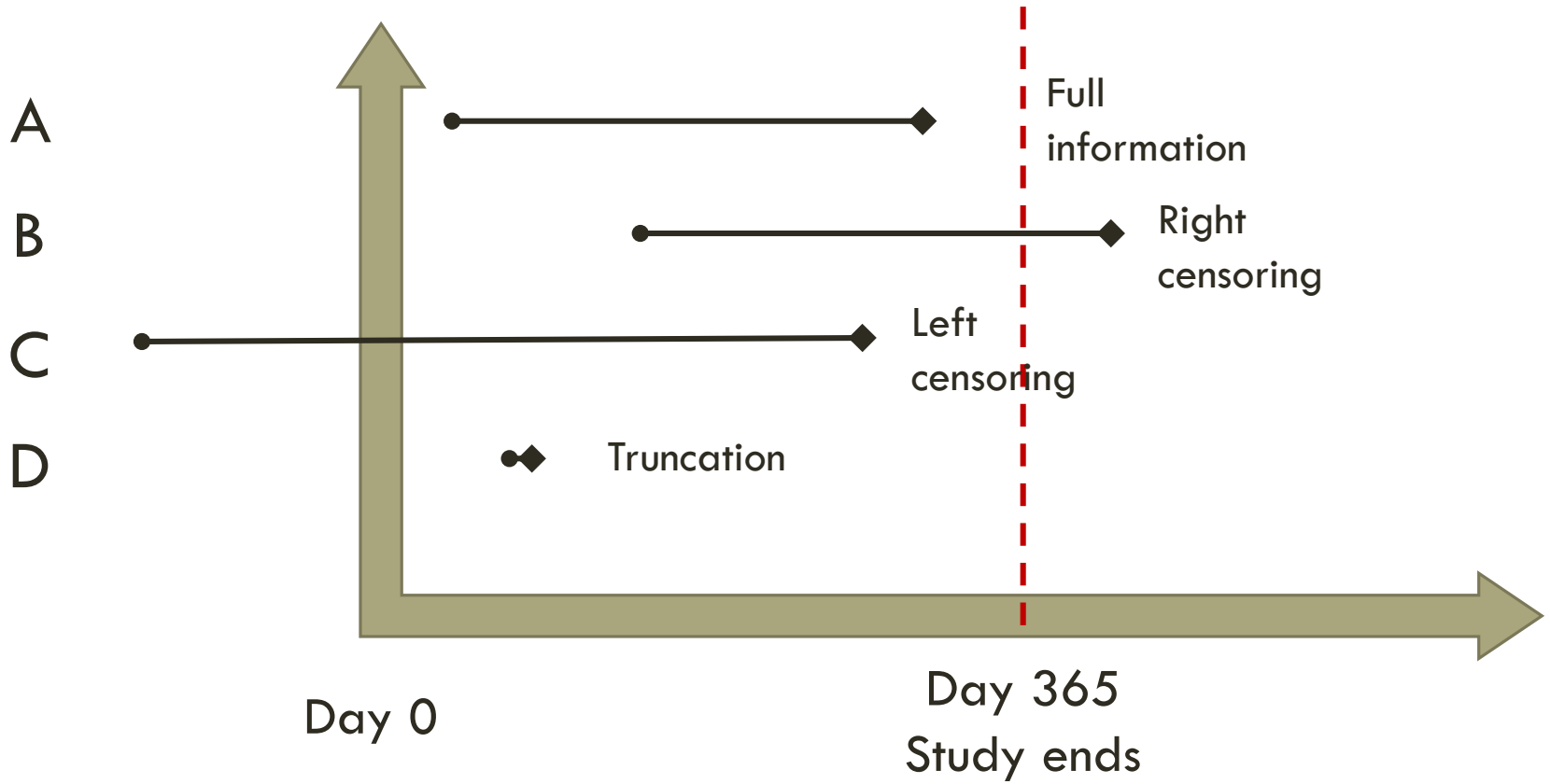
   The traditional "solution" is to include as many variables as possible in the model. *Why is this a problem?*

# MORE TERMINOLOGY TO BE FAMILIAR WITH

The following terms are usually paired, some of which we'll return to in later slides:

- **Unit versus Item Non-Response.** Generally used in the context of surveys or designed studies. *Unit* non-response is when a particular question/variable is missing; *Item* non-response is failure to obtain any data from a participant.
- **Truncation versus Censoring.** Truncation is when a participant is not included because of the values of certain variables; Censoring is when a variable is not fully known/specified.
- **Ignorable versus Non-Ignorable missingness.** MCAR and MAR are usually called "ignorable" because, as we'll discuss in the next section, there are methods for handling without changing our model.

# TRUNCATION AND CENSORING




We want to understand time between purchasing a piece of machinery and its first breakdown.

We start recording at Day 0, and after a year (Day 365) we want to draw conclusions by looking at internal maintenance records.

Machinery that breaks down within the first month is returned to the manufacturer, and not in our records.

# HANDLING OF MISSING DATA

# OUR PROCEDURE FOR HANDLING MISSING DATA

1. Perform EDA
2. Make assumptions which are reasonable given the data and our subject matter expertise
3. Document our assumptions
4. Perform sensitivity analysis

# SPECTRUM OF METHODS

| | |
|---|---|
| Discarding data | • Complete case<br>• Available case |
| Single imputation | • Mean and regression imputation<br>• Create a dummy variable |
| Model-based & multiple imputation | • Maximum likelihood method<br>• Multiple imputation |

# DISCARDING DATA

For complete case methods, we discard any observation that has even one missing value, e.g. isn't entirely complete.

**Pros**: Very simple.

**Cons**:

- Can produce biased estimates
- May throw out important info in an observation (i.e. inefficient)

| Gender | Ice Cream |
|---|---|
| F | Chocolate |
| F | Mint |
| F | Chocolate |
| F | Chocolate |
| M | Chocolate |
| ~~M~~ | ~~?? (Mint)~~ |
| M | Mint |
| ~~M~~ | ~~?? (Mint)~~ |

# SINGLE IMPUTATION OFTEN BIAS IN CASES OF MAR

*Mean imputation* replaces a missing observations with the mean of the non-missing values of the same variable.

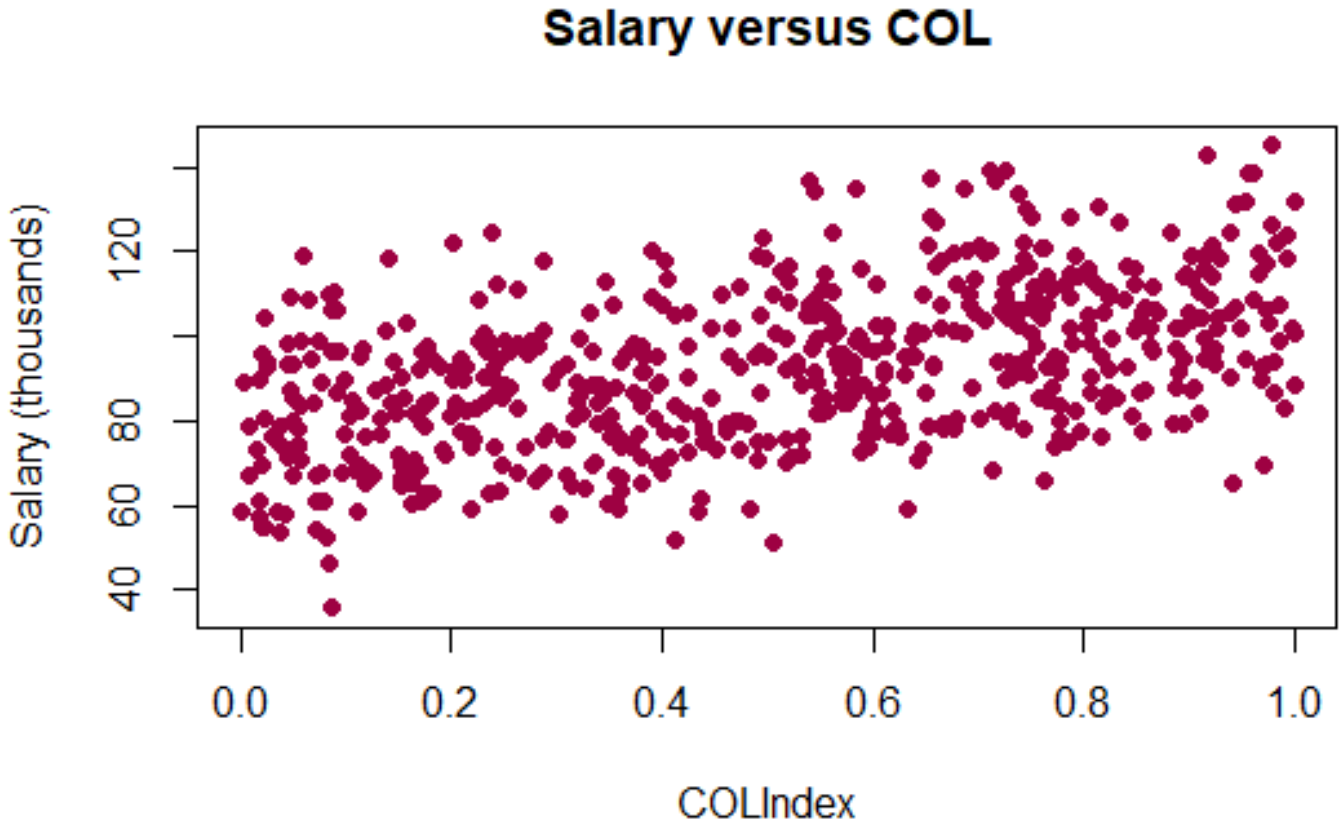
- Why is this a problem if the data are MAR or MNAR?
- Let's explore why this is a problem *even if the data are MCAR*…

| Gender | Ice Cream |
|---|---|
| F | 3.5 |
| F | 4 |
| F | 3 |
| F | 5 |
| M | 3 |
| M | ~~2~~ 3.4 |
| M | 2 |
| M | ~~1~~ 3.4 |

# SINGLE IMPUTATION METHODS UNDERSTATE UNCERTAINTY

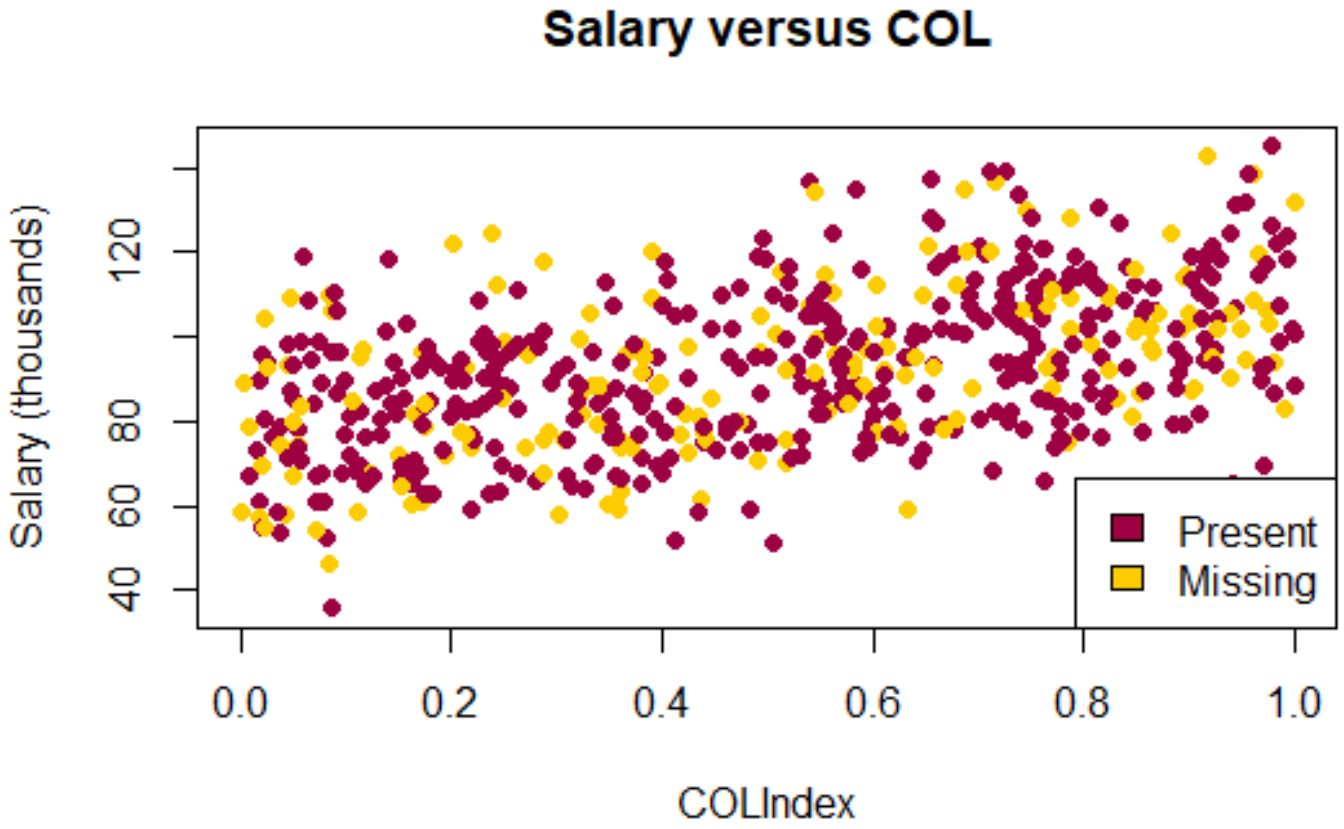
- Let's simulate salary data as a function of **cost of living** and **years in grad school**

```
Salary ~ 60 +
30•COLIndex +
20•yearsGradSchool
```



# SINGLE IMPUTATION METHODS UNDERSTATE UNCERTAINTY

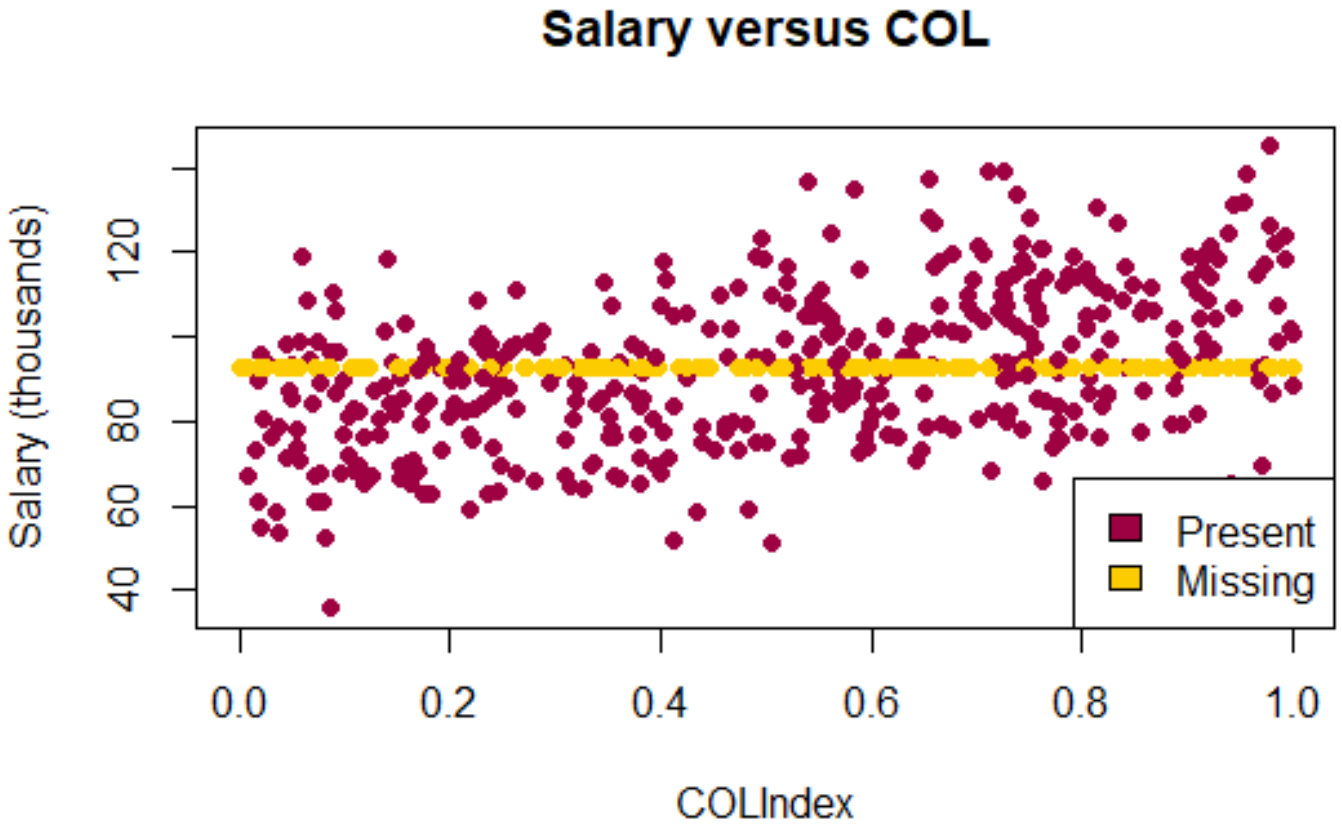
Next, set `Salary` records to `NA` with a independent probability of 30%



# SINGLE IMPUTATION METHODS UNDERSTATE UNCERTAINTY

Perform Mean Imputation.

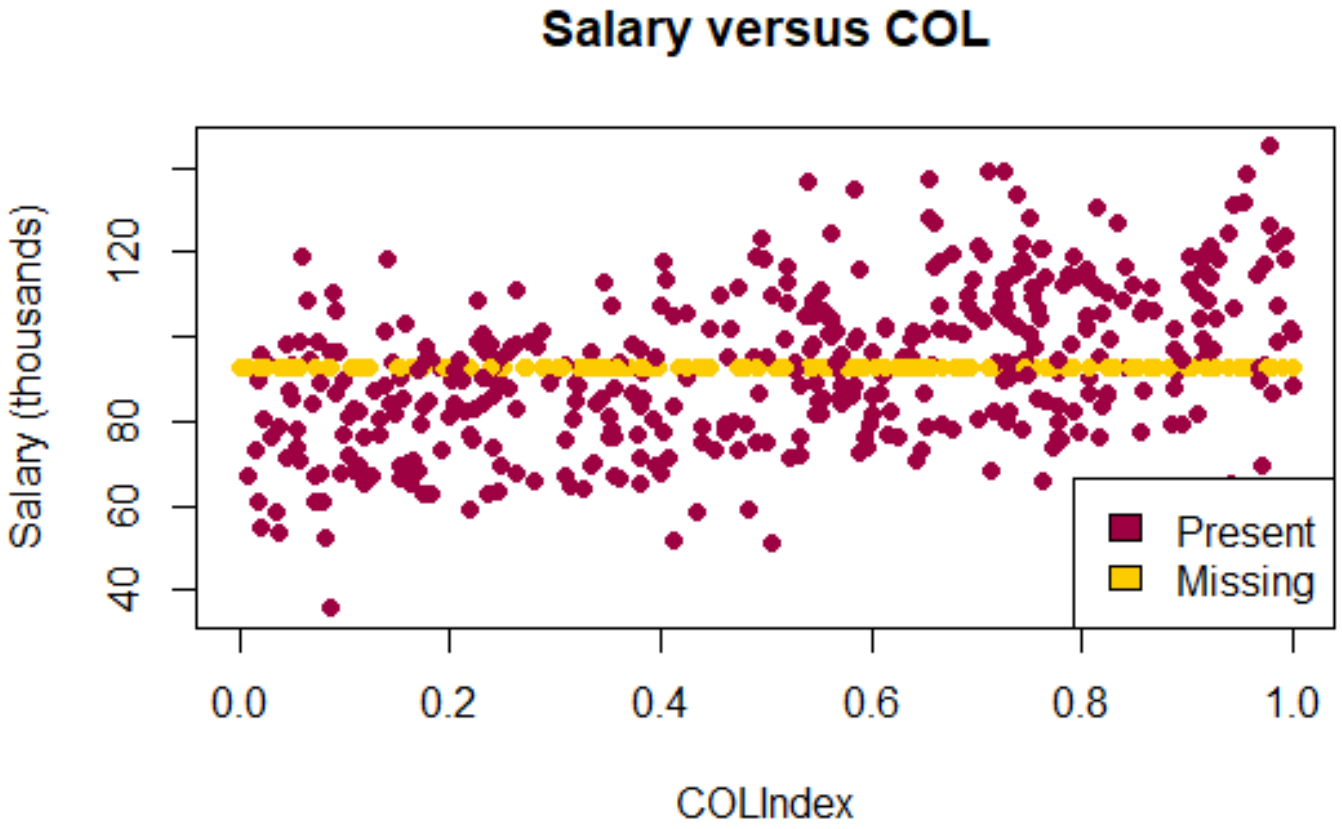
Is the relationship between `Salary` and `COLIndex` unchanged?



# SINGLE IMPUTATION METHODS UNDERSTATE UNCERTAINTY

Lastly, regress `Salary` on `COLIndex` and `yearsGradSchool`.

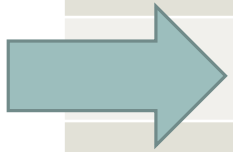
Does `lm` "know" that some of the data is imputed?



# WHY MENTION IMPUTATION IF IT'S SO TERRIBLE?

Assume that our ice cream example had 1,000 mostly complete variable.

We don't want to lose all the information in mostly complete row due to one missing item.



| Gender | Ice Cream | $X_3$ | … | $X_{1000}$ |
|---|---|---|---|---|
| F | 3.5 | T | … | 4.2 |
| F | 4 | F | … | 5.3 |
| F | 3 | T | … | 1.1 |
| F | 5 | T | … | -0.1 |
| M | 3 | F | … | 20.3 |
| M | ~~2~~ | T | … | -12.4 |
| … | … | … | … | … |
| M | ~~1~~ | F | … | 19.2 |

# DOWNSIDES OF SINGLE IMPUTATION

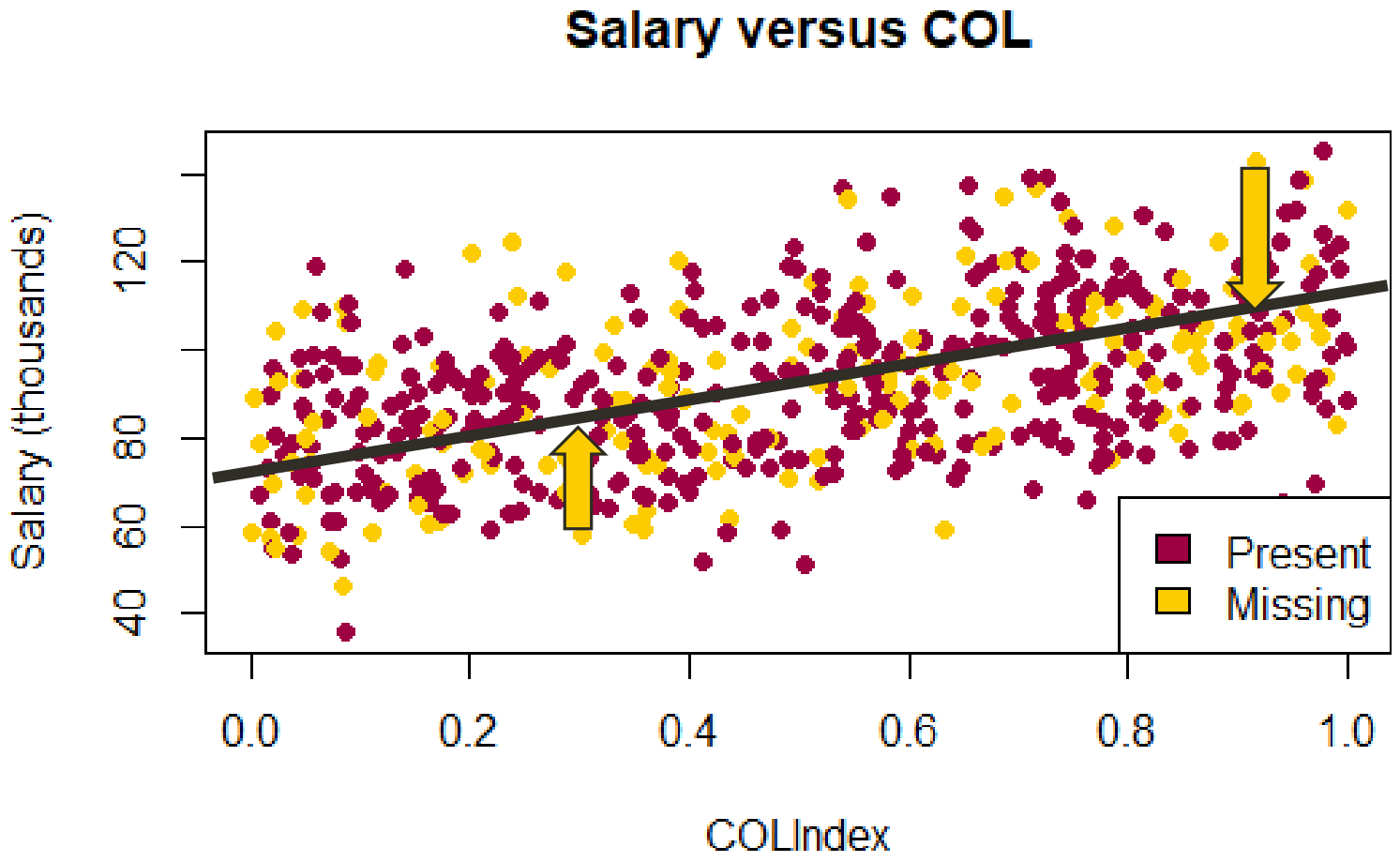
| Method | Description | Downsides |
|---|---|---|
| Mean and regression imputation | Use a sensible approach to impute a single value | Understate uncertainty |
| Dummy variable | Create a new category for all missing observations | May group together wildly different values |
| Matching/hot-deck methods | Find a similar record (e.g. nearest neighbor) and use its value for the missing cell | Over-use of certain neighbors/donors |

# MULTIPLE IMPUTATION CARRIES FORWARD UNCERTAINTY

- The goal of multiple imputation is to carry through uncertainty about the imputed values to our final inferences
- Methodology:
  - Add variation/uncertainty to the imputation
  - Perform analysis on the imputed data set
  - Repeat this many times
  - Summarize the results to produce parameter estimates, standard errors, and other inferences

# MULTIPLE IMPUTATION CARRIES FORWARD UNCERTAINTY

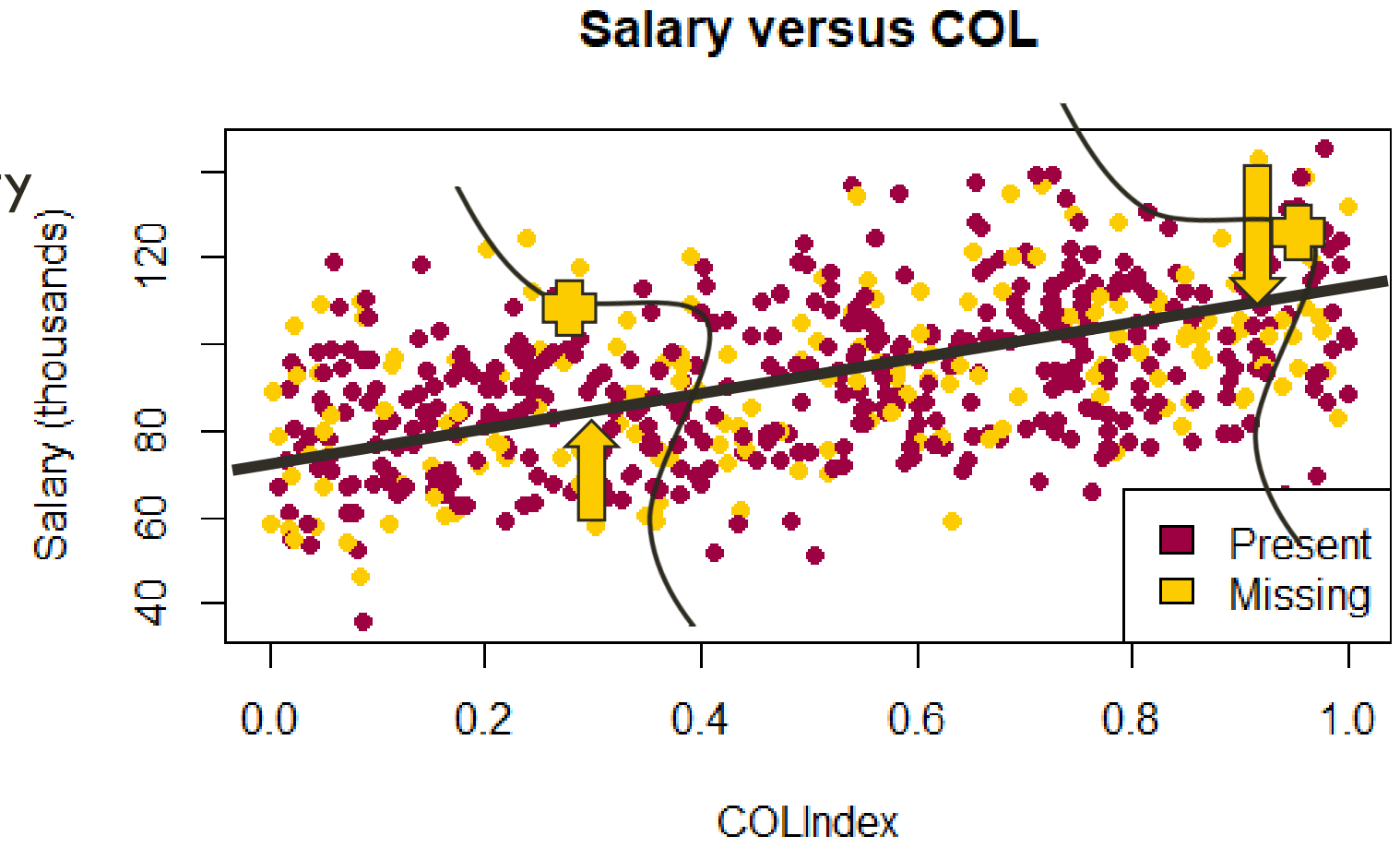
- This is where single linear regression imputation would stop



# MULTIPLE IMPUTATION CARRIES FORWARD UNCERTAINTY

MI:

- adds uncertainty
- samples an imputed value
- then performs complete data analysis
- repeats this many times



# MAXIMUM LIKELIHOOD METHOD INTEGRATES THE MISSINGNESS AWAY

- Recall that *Maximum Likelihood methods* attempt to find a set of parameters that maximize the probability of having seen the observed data

General procedure for ML:

- Specify the full likelihood function $\mathbf{L(\theta \mid X) = \prod f_i(X \mid \theta)}$, and then find the values for $\theta$ that **maximize the likelihood**

# MAXIMUM LIKELIHOOD METHOD INTEGRATES THE MISSINGNESS AWAY

How to handle missing data in ML context?

1. Group the data by identical missingness patterns
2. Factor the likelihood into groups of identical likelihood
3. For each group, integrate out the missing variables
4. Maximize the resulting likelihood function

# MAXIMUM LIKELIHOOD METHOD INTEGRATES THE MISSINGNESS AWAY

Suppose we have two variables, $X_1$ and $X_2$, only $X_1$ has any missingness.

Let M = set of rows in which $X_1$ is missing; for i in M,

$$g_i(x_1, x_2 \mid \theta) = \int f_i(x_1, x_2 \mid \theta)\, dx_1$$

Then

$$L(\theta \mid X) = \prod f_i(X \mid \theta) = \prod g_M(x_2 \mid \theta) \prod f_{Mc}(x_1, x_2 \mid \theta)$$

# DOWNSIDES OF MI AND ML METHODS

## Multiple Imputation

- Requires many decisions
- Can be computationally intensive
- Assumptions of imputation model (e.g. linearity) may not match analysis model
- Not all test statistics can be aggregated (e.g. p-value)

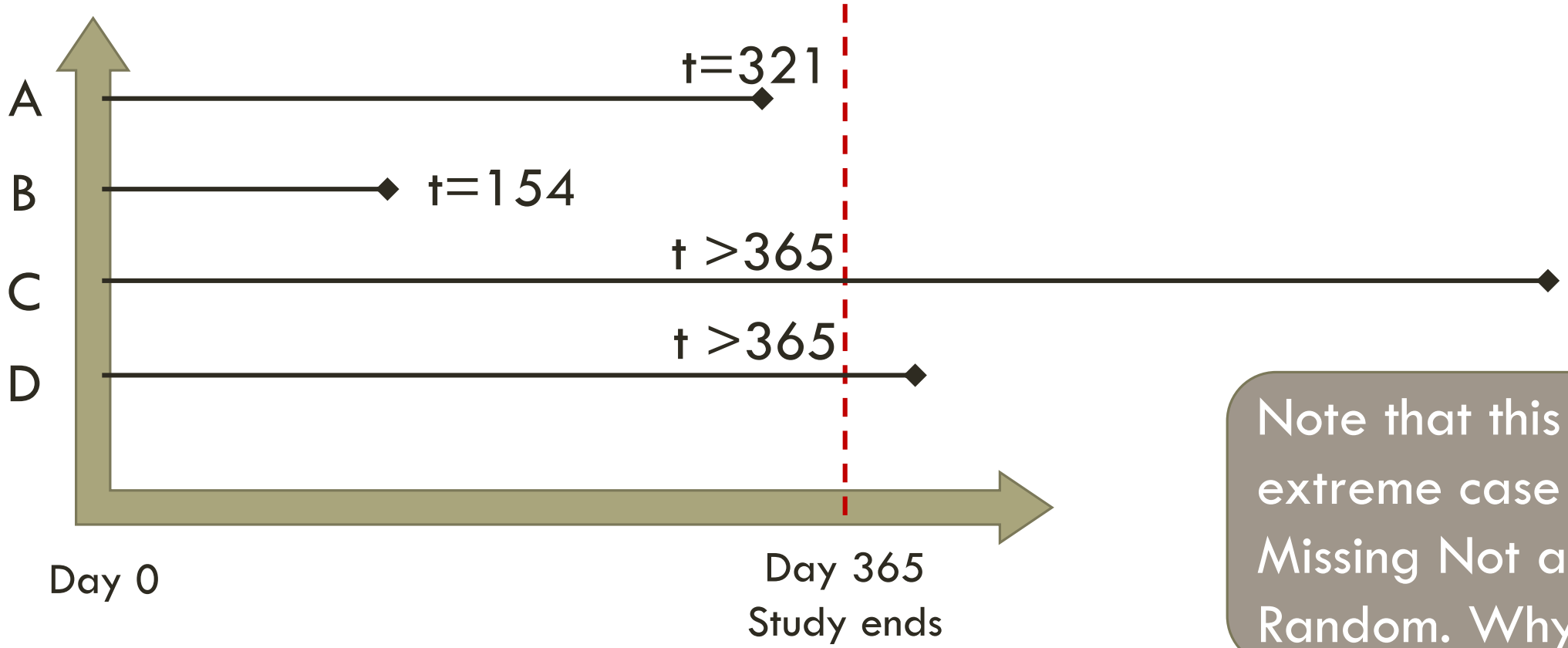
## Maximum Likelihood

- Need independence to easily factor the likelihood
- Need be linear model to easily integrate out a variable

# TO RECAP SO FAR

| | | |
|---|---|---|
| If data are MCAR*… | Complete case analysis or MI/ML methods are appropriate | *But it rarely is… best you can hope for is cannot reject a null hypo in MCAR test |
| If data are MAR… | MI/ML methods are appropriate | But in the era of big data, need to be smarter than just throwing all variables in to be safe |
| If data are MNAR… | Are we stuck? | No – but need to adapt the model to the missingness pattern |

# TOBIT MODELS FOR CENSORED REGRESSION




Note that this is an extreme case of Missing Not at Random. Why?

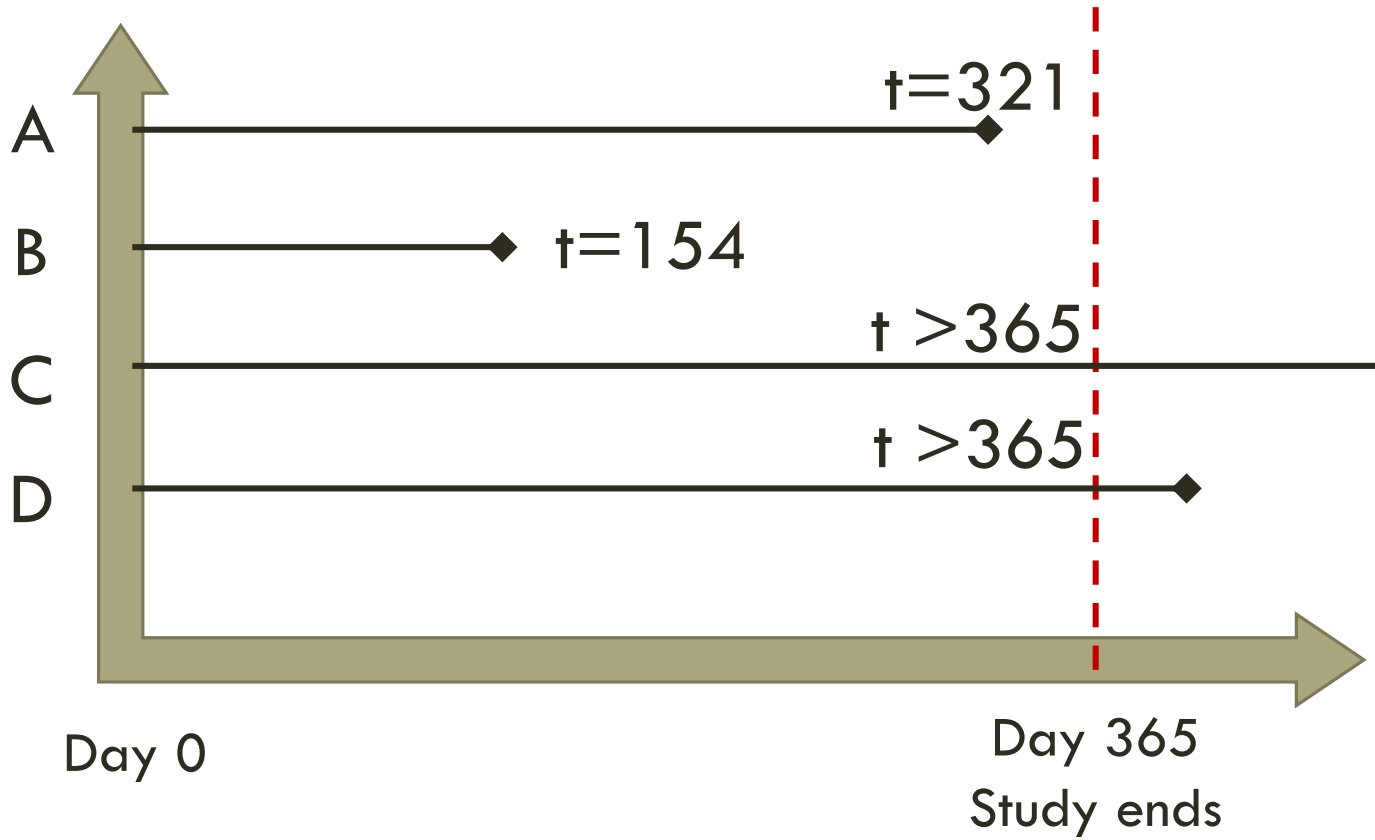
# TOBIT MODELS FOR CENSORED REGRESSION

Not an imputation method at all!

`censReg` in R

- As in the ML imputation method, Tobit models work by factoring the likelihood into missingness patterns, and adapting the Likelihood function
- The precise form will depend on the regression model selected
- Then as with all ML methods, find the parameter values that maximize the likelihood

# TOBIT MODELS FOR CENSORED REGRESSION




The likelihood would multiply the probability (pdf/pmf) of the durations of participants A and B by the probability of durations being greater than or equal to 365 (1minus the cdf/cmf) for participants C and D

# TOBIT MODELS FOR CENSORED REGRESSION

Suppose we want to model $Y = X\beta$, but Y has some censored values as on the previous slide.

As usual, let $\Phi$ represent the cdf for a standard normal, and $\phi$ represent its pdf.

Then

$$L(\beta \mid X,Y) = \prod f_i(Y \mid \beta,X) = \prod f_{A,B}(Y \mid \beta,X) \prod f_{C,D}(Y \mid \beta,X)$$

$$= \prod_{A,B} \frac{1}{\sigma} \phi\left(\frac{Y - X\beta}{\sigma}\right) \prod_{C,D} \frac{1}{\sigma}\left(1 - \Phi\left(\frac{365 - X\beta}{\sigma}\right)\right)$$

# KNOW WHAT YOUR ALGORITHM IS DOING!

What is the default behavior in `lm`?

`na.omit` returns the object with incomplete cases removed.

What about `rpart`?

`na.action` the default action deletes all observations for which y is missing, but keeps those in which one or more predictors are missing.

My experience/bias – more rework or flat-out poor inference has been generated by misunderstanding default missingness behavior of algorithms than any subtle differences in imputation methodologies

# VALIDATION AND SENSITIVITY ANALYSIS

**Over-imputation:** the missing data equivalent of cross-validation

**Methodology:** Treat some observed data as if it were missing, and determine if all your assumptions provide inferences that are consistent with the known, held-out data

**Technology:** **`Amelia`** in R

# OUR PROCEDURE FOR HANDLING MISSING DATA

1. Perform EDA
2. Make assumptions which are reasonable given the data and our subject matter expertise
3. Document our assumptions
4. Perform sensitivity analysis

# APPENDIX

# MISSINGNESS AS IT RELATES TO SOME CLASS PROJECTS

- Kadane Fingerprint Survey
- CivicScience
- Fox Chapel Schools
- RRP

# SPECIAL CASES

1. Time series missingness
2. Bayesian missingness
3. Sparseness

# WHAT IS MISSING DATA?

Suppose that I have taken a cruise, and I leave the following survey question blank:

How would you rate the food quality on your most recent cruise?

- Excellent
- Good
- Fair
- Poor

# WHAT IS MISSING DATA?

Suppose that I have not taken a cruise, and I leave the following survey question blank:

How would you rate the food quality on your most recent cruise?

- Excellent
- Good
- Fair
- Poor
- I have never taken a cruise or can't recall the food

# WHAT IS MISSING DATA?

Suppose that I have not taken a cruise, and I leave the following survey question blank:

How would you rate the food quality on your most recent cruise on a scale from 0 to 100?

- 50
- 0
- -999 (if no cell validation)
- blank