

Traps of Missing Values

Hua Shi

Modeling & Analytics

Customer Marketing





“Why don’t you just replace them with zero?”

“Why not any other value?”

Random Missing v.s. Structural Missing

- Data collection errors
- Incomplete customer responses
- System/Measurement failures
- Revision of data collection scope
- Not applicable

x	y
1	.
.	4
.	9
4	.
5	25
.	.
7	.
8	64

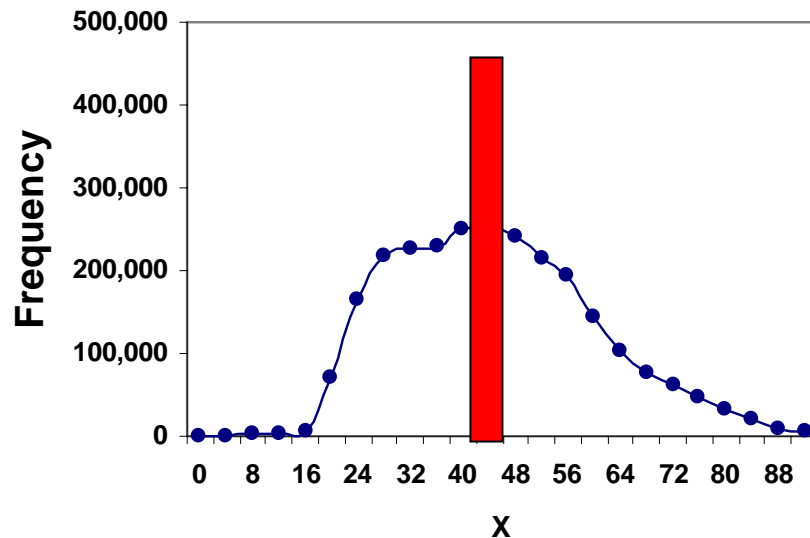
x	y
0	.
0	.
0	.
0	.
1	361
3	81
3	324
4	169

What to Do with Missing Values?

- Do nothing – treat it as a separate category
- Reject observations with missing value
- Reject variables with missing value
- Split the data and build separate models

Impute a Single Value

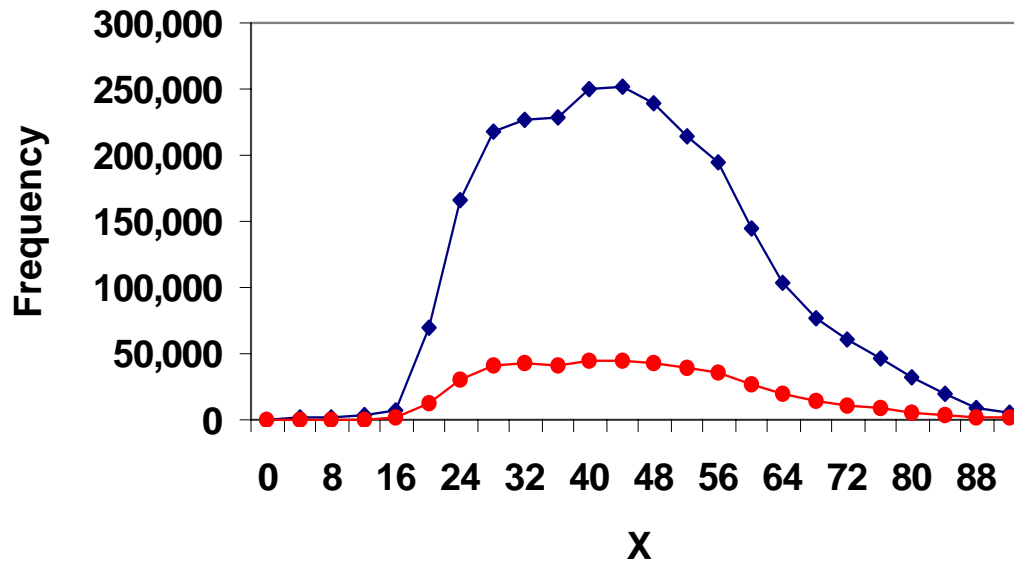
If $x = \text{missing}$, $x = 44$ (Mean, Median, Mid-range, Mode)



- Simple
- but affects the variable's sample distribution
- Always create missing value indicator

Impute Multiple Values - Distribution

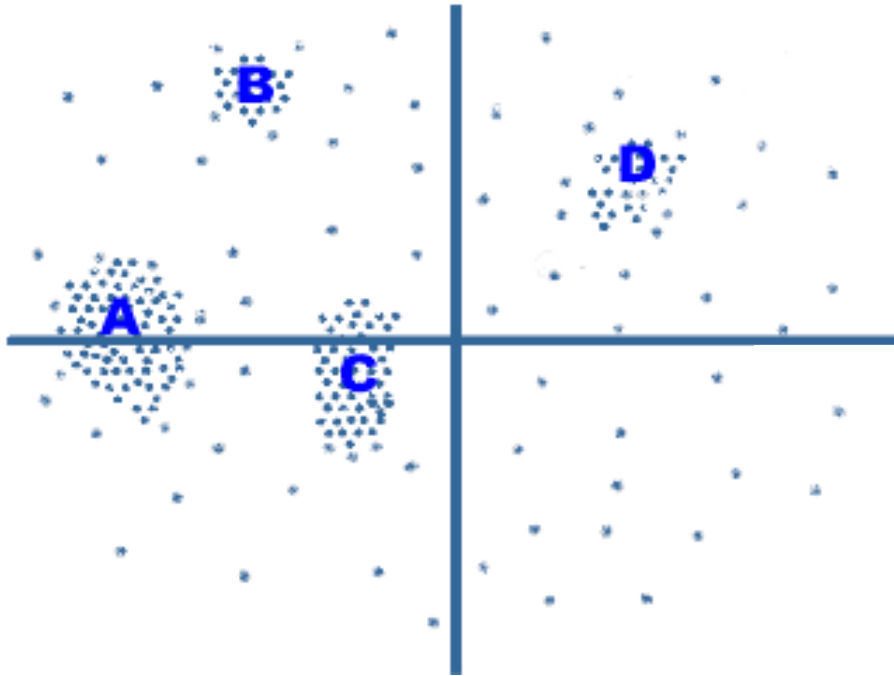
Randomly impute missing values to mimic distribution of non-missing observations



- More difficult to do
- Do not change the distribution of data
- Not necessarily more accurate

Impute by Cluster Analysis

If x = missing, \hat{x} = cluster mean.

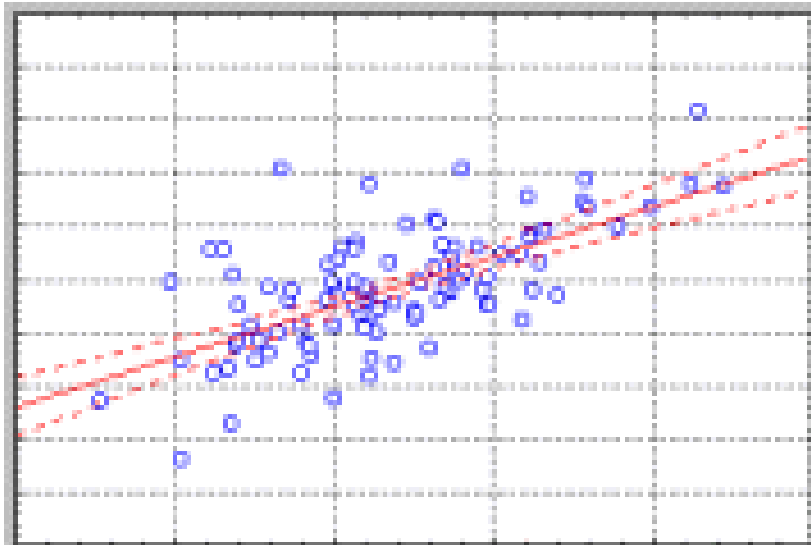


- Can be more accurate
- Need to impute other inputs first
- Much more work.

Impute by Regression

Treat the variable with missing value as target, the remaining variables as predictors.

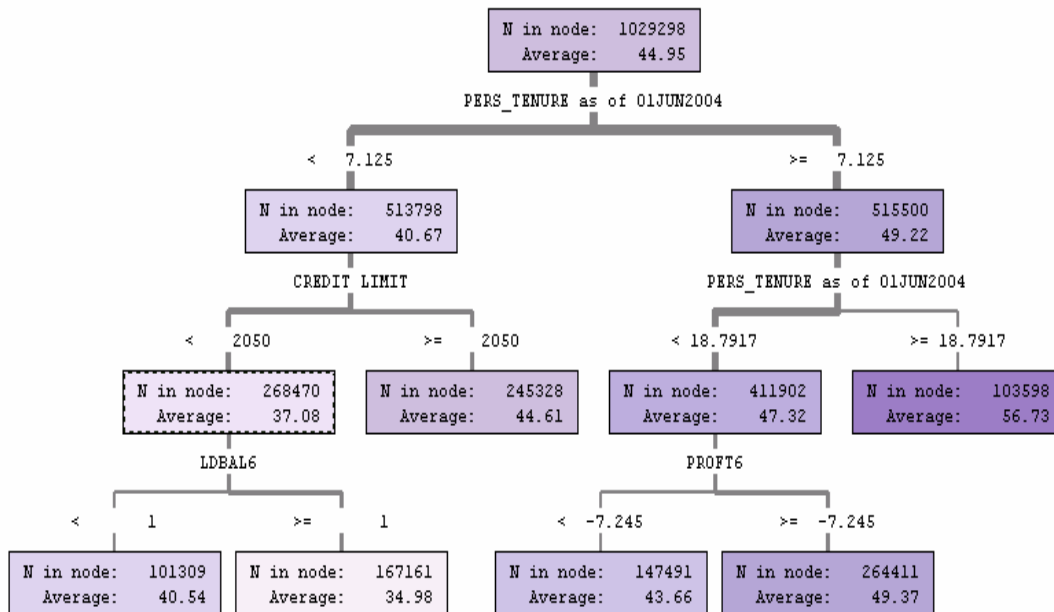
If x = missing, then $x = b_0 + b_1 * x_1 + b_2 * x_2 + \dots$



- Can be more accurate.
- Need to impute other inputs first.
- Much more work.

Tree Imputation

If x = missing, x = node average

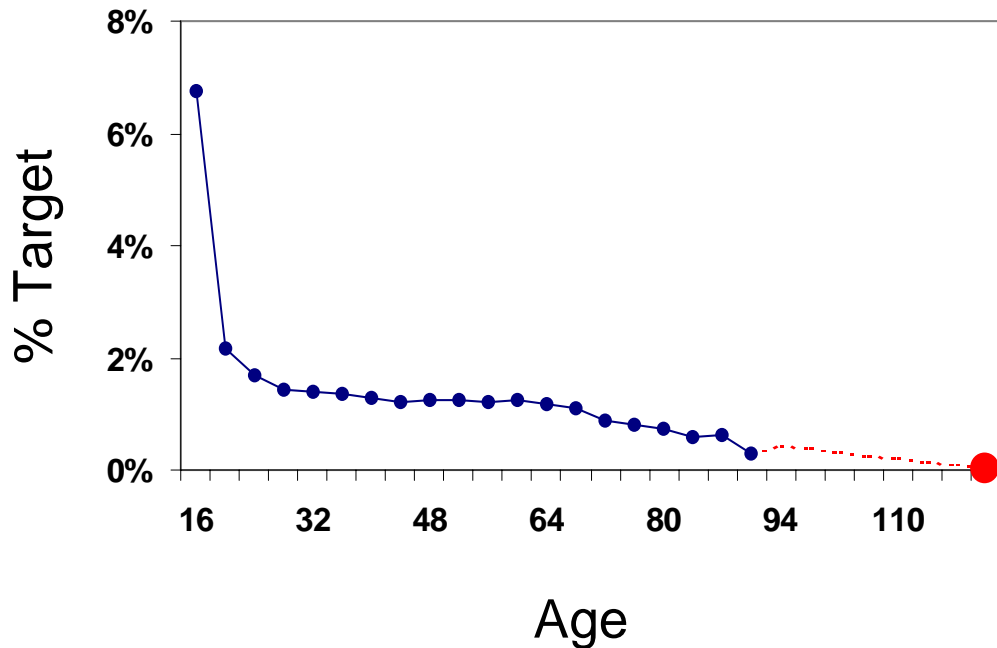


- Can be more accurate
- No need to impute other inputs.
- Impute discrete values

Impute Based on Target Behavior

Missing age group contain 0.04% targets

if age = missing, then age = 120.



- Not necessarily more accurate
- Helpful for prediction
- Again, create missing value indicator

Missing Values in Computation

x	2	4	-6	•	8	10
---	---	---	----	---	---	----

- Missing is the smallest value

“x < 5” finds 2, 4, -6, •

- Cause missing in direct math computation

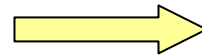
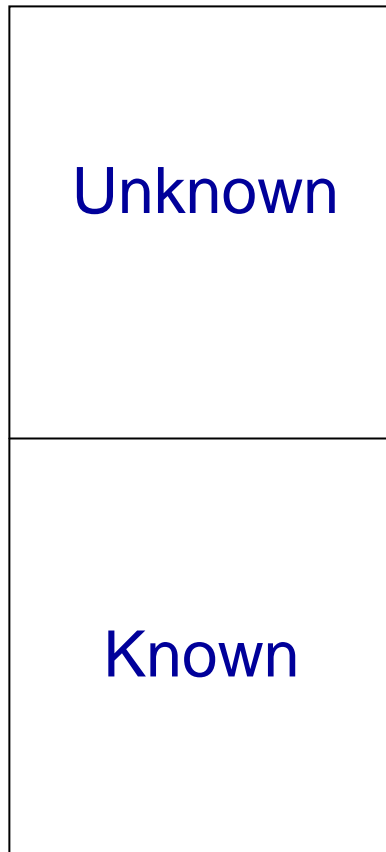
2 + 4 + (-6) + • + 8 + 10 = • ; log(2 + •) = • ;

- Ignored in summarizing functions

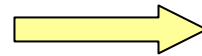
sum(..) = 18 ; mean(..)=18 / 5=3.6 ; min(..) = -6 ;

- Pairwise v.s. listwise deletion

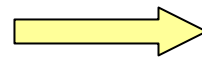
One Missing, Multiple Meanings



Couldn't reach



Refused to answer



Question not applicable

Quiz: in SAS, up to how many different missing values can you use for one single variable?

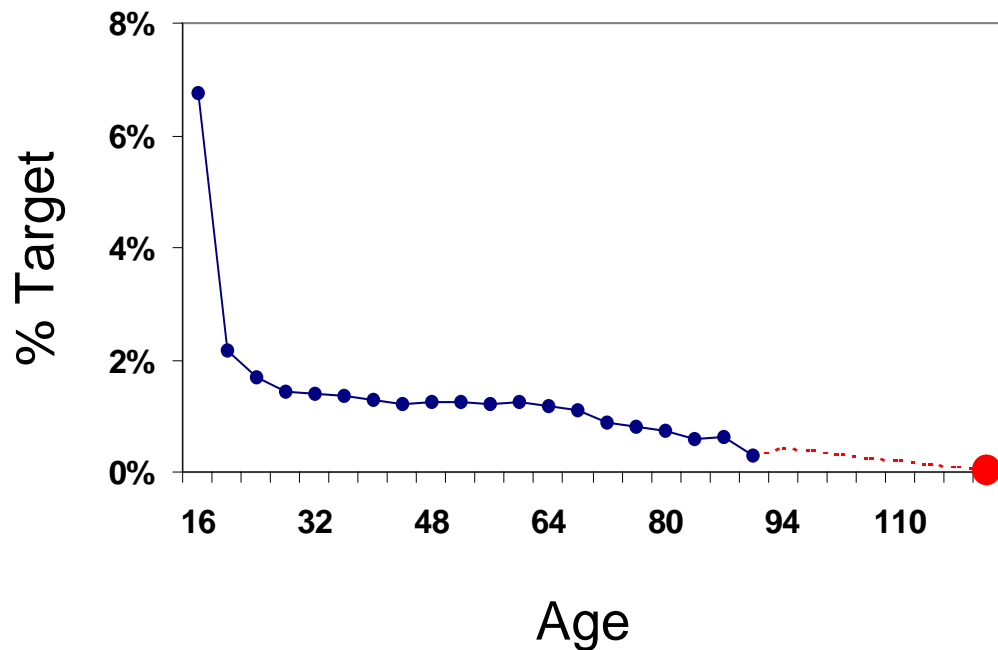
28

Missing Disguised as Not Missing

- Value 999
- Partial dates, defaulted to day 1 the month
- Impossible values – negative age
- Outliers

Dangerously Missing

Missing age group have very few targets. Why?



- Age was updated
- When target behavior was exhibited, it was filled in
- It remained missing for non-targets.
- Missing age contains future information.

In Conclusion ...

Missing value handling is more of an art than science.

There is no single silver bullet.

“The artist never entirely knows.

We guess.

We may be wrong,

But we take leap (of faith) after leap

in the dark.”



Agnes De Mille