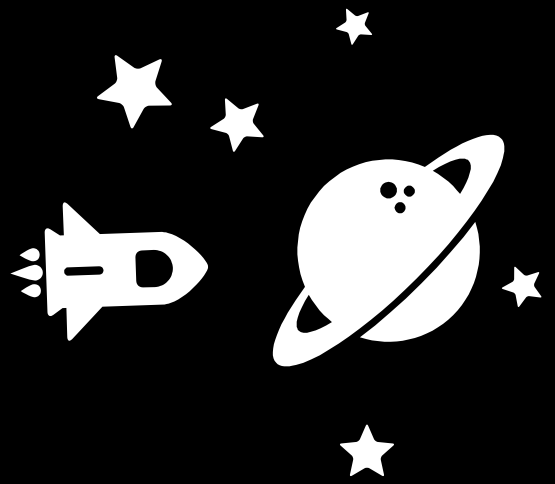
Missing Data Handling

Presentation by Berk Sudan

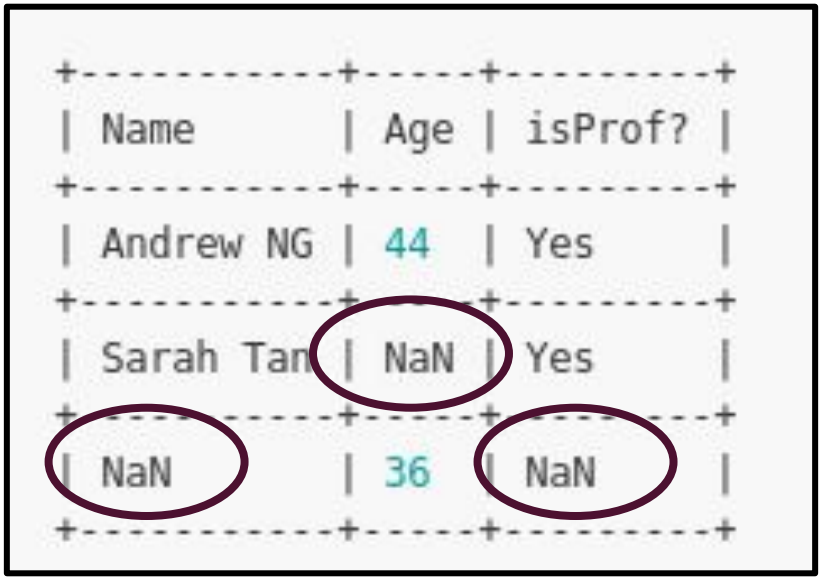
b	25.5	0.375	u	g	m	v	0.25	t	t	3	f	g	260	15108	+
b	19.42	6.5	u	g	w	h	1.46	t	t	7	f	g	80	2954	+
b	35.17	25.125	u	g	x	h	1.625	t	t	1	t	g	515	500	+
b	32.33	7.5	u	g	e	bb	1.585	t	f	0	t	s	420	0	-
b	34.83	4	u	g	d	bb	12.5	t	f	0	t	g	?	0	-
a	38.58	5	u	g	cc	v	13.5	t	f	0	t	g	980	0	-
b	44.25	0.5	u	g	m	v	10.75	t	f	0	f	s	400	0	-
b	44.83	7	y	p	c	v	1.625	f	f	0	f	g	160	2	-
b	20.67	5.29	u	g	q	v	0.375	t	t	1	f	g	160	0	-
b	34.08	6.5	u	g	aa	v	0.125	t	f	0	t	g	443	0	-
a	19.17	0.585	y	p	aa	v	0.585	t	f	0	t	g	160	0	-
b	21.67	1.165	y	p	k	v	2.5	t	t	1	f	g	180	20	-
b	21.5	9.75	u	g	c	v	0.25	t	f	0	f	g	140	0	-
b	49.58	19	u	g	ff	ff	0	t	t	1	f	g	94	0	-
a	27.67	1.5	u	g	m	v	2	t	f	0	f	s	368	0	-
b	39.83	0.5	u	g	m	v	0.25	t	f	0	f	s	288	0	-
a	?	3.5	u	g	d	v	3	t	f	0	t	g	300	0	-
b	27.25	0.625	u	g	aa	v	0.455	t	f	0	t	g	200	0	-
b	37.17	4	u	g	c	bb	5	t	f	0	t	s	280	0	-
b	?	0.375	u	g	d	v	0.875	t	f	0	t	s	928	0	-
b	25.67	2.21	y	p	aa	v	4	t	f	0	f	g	188	0	-

Ref: *https://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.data*

Missing at random or not?



Missing at Random

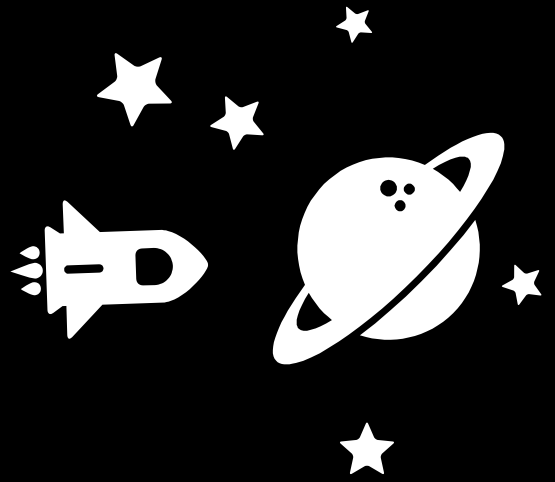


Name	Age	isProf?
Andrew NG	44	Yes
Sarah Tan	NaN	Yes
NaN	36	NaN

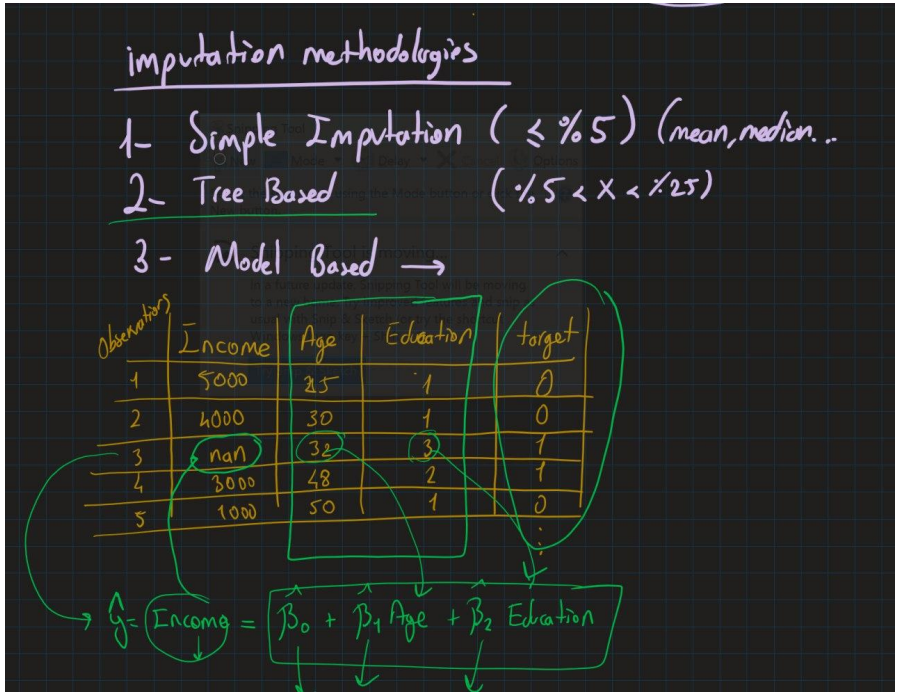
Missing at not Random

Name	Age	Type of House
Andrew NG	44	A
Sarah Tan	29	B
Lex Fridman	36	NaN

Missing Value Imputation



Methodologies



Ref: *Çağlar Subaşı, 2020 Kodluyoruz Applied Data Science-101 Bootcamp, Lecture 3*

Simple Imputation

- For **numerical** values:
  - If has **many outliers**: Use MEDIAN
  - If has **fewer outliers**: Use MEAN
- For **categorical** values:
  - **Mode** can be used.
  - Alternatively, **stratified sampling**

Ref: *Çağlar Subaşı, 2020 Kodluyoruz Applied Data Science-101 Bootcamp, Lecture 3*

Tree-based

- **1st feature:** The feature which has missing value(s)
- **2nd feature:** Target feature
- So, use instances with **missing values** as test-set, predict the missing value!

Ref: *Çağlar Subaşı, 2020 Kodluyoruz Applied Data Science-101 Bootcamp, Lecture 3*

Model-Based

- Use **all features** to predict missing value!
- Mark the **variable** as **target**

Ref: *Çağlar Subaşı, 2020 Kodluyoruz Applied Data Science-101 Bootcamp, Lecture 3*

What If Many Missing Values?

- Ensure missing values are occurred **randomly**.
- Ensure missing values means **nothing (**or **anything)**.
- **Delete** the feature/variable!

Ref: *Çağlar Subaşı, 2020 Kodluyoruz Applied Data Science-101 Bootcamp, Lecture 3*

End of Presentation

Presented by Berk Sudan