Handling Missing Data

Ravikiran Srinivasulu
SOFTWARE CONSULTANT

ravikirans.com | go.ravikirans.com/YouTube

Agenda

We cannot control for missing data

Approach the problem in a systematic way

Understand the types of missingness

In this module:

- Listwise deletion
- Custom Substitution
- Single Imputation methods
- MICE

Complete dataset

Reasons Why Data Is Missing

Reasons of Missing Data

Missing Completely at Random (MCAR)

Missing at Random (MAR)

Missing Not at Random (MNAR)

Missing Completely at Random (MCAR)

Country	Degree
United States	Bachelors
Cambodia	Masters
India	Preschool
Mexico	Bachelors
?	Masters
Germany	Doctorate
?	Masters
England	9th
Italy	11th
Columbia	HS-grad

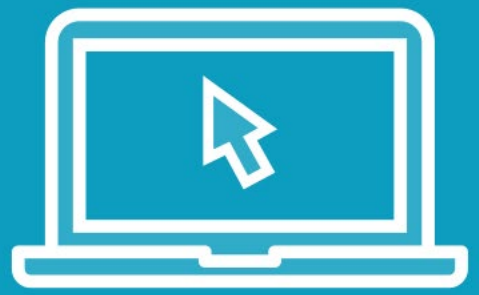
Missing at Random (MAR)

Age	Gender
35	Male
25	Male
32	Female
?	Female
?	Female
30	Male
?	Male
55	Male
?	Female
35	Male

Missing Not at Random (MNAR)

Occupation
Priv-house-serv
Handlers-cleaners
Armed-Forces
?
?
Farming-fishing
?
Other-service
Exec-managerial
Sales

Demo



Listwise Deletion

Problems in Deleting Rows

Problems in Deleting Rows

Deleting records introduces bias

Data missing at prediction time

Loss of data impacts high variance models

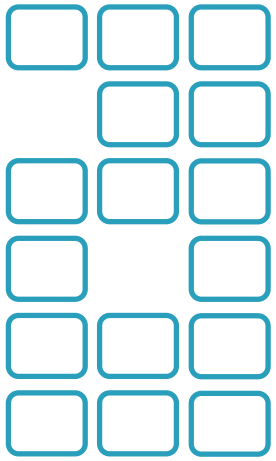
Problems in Deleting Rows

Deleting records introduces bias

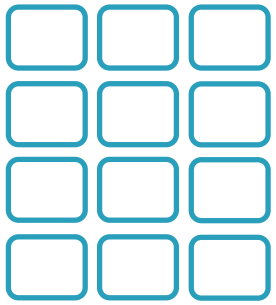
Listwise deletion works only if the assumption is MCAR

- **MAR – Women not revealing their age**
- **MNAR – High-salaried people not disclosing their incomes**

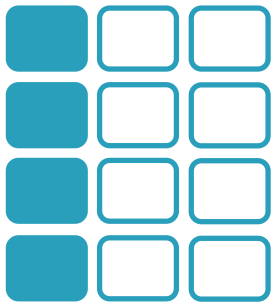
What if Data Is Missing at Prediction Time?

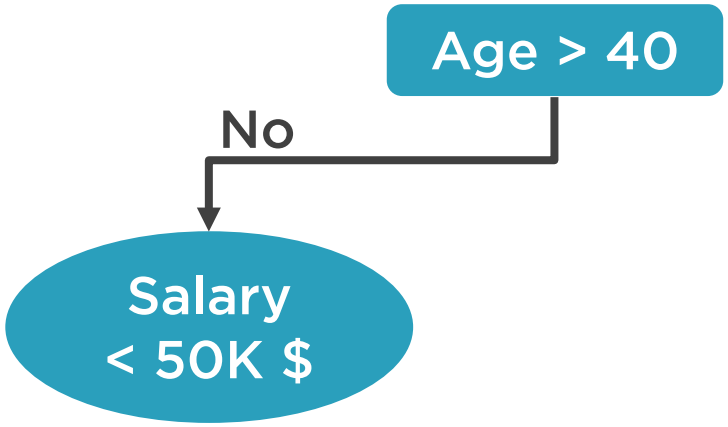


What if Data Is Missing at Prediction Time?

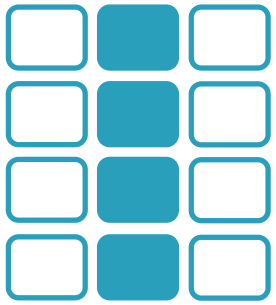


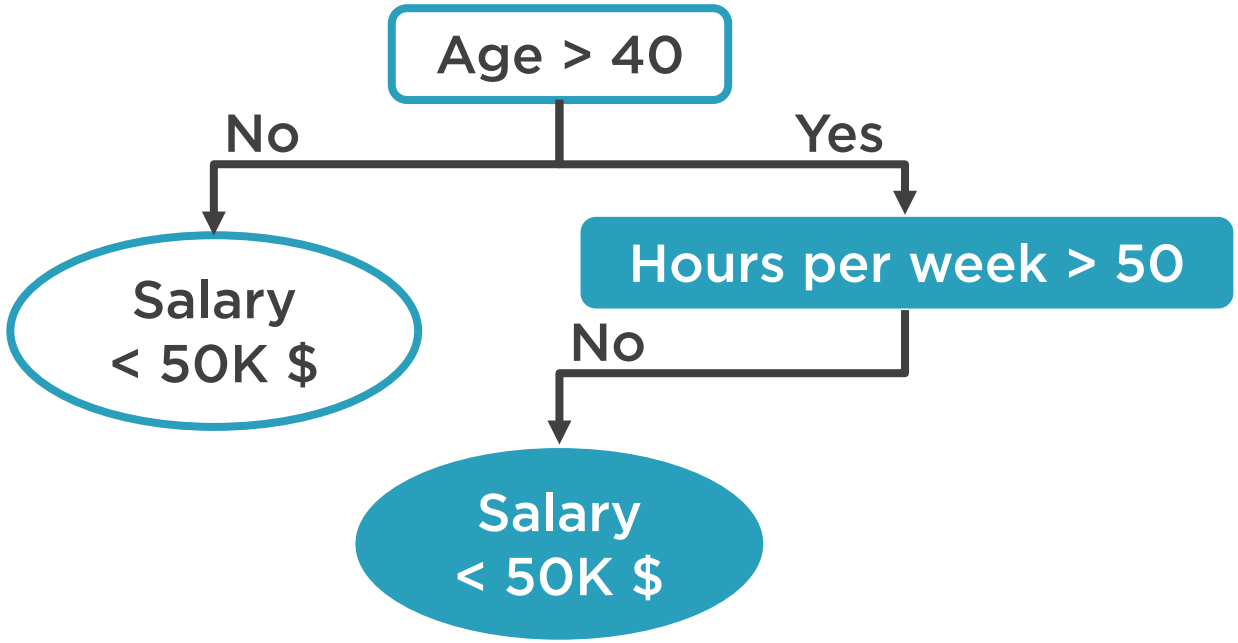
What if Data Is Missing at Prediction Time?



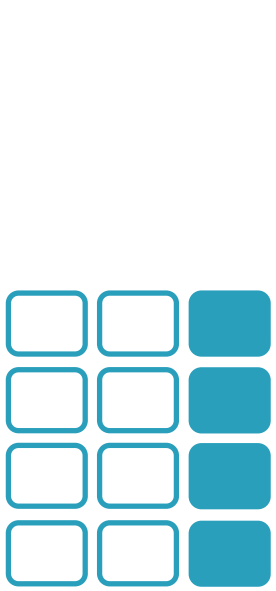


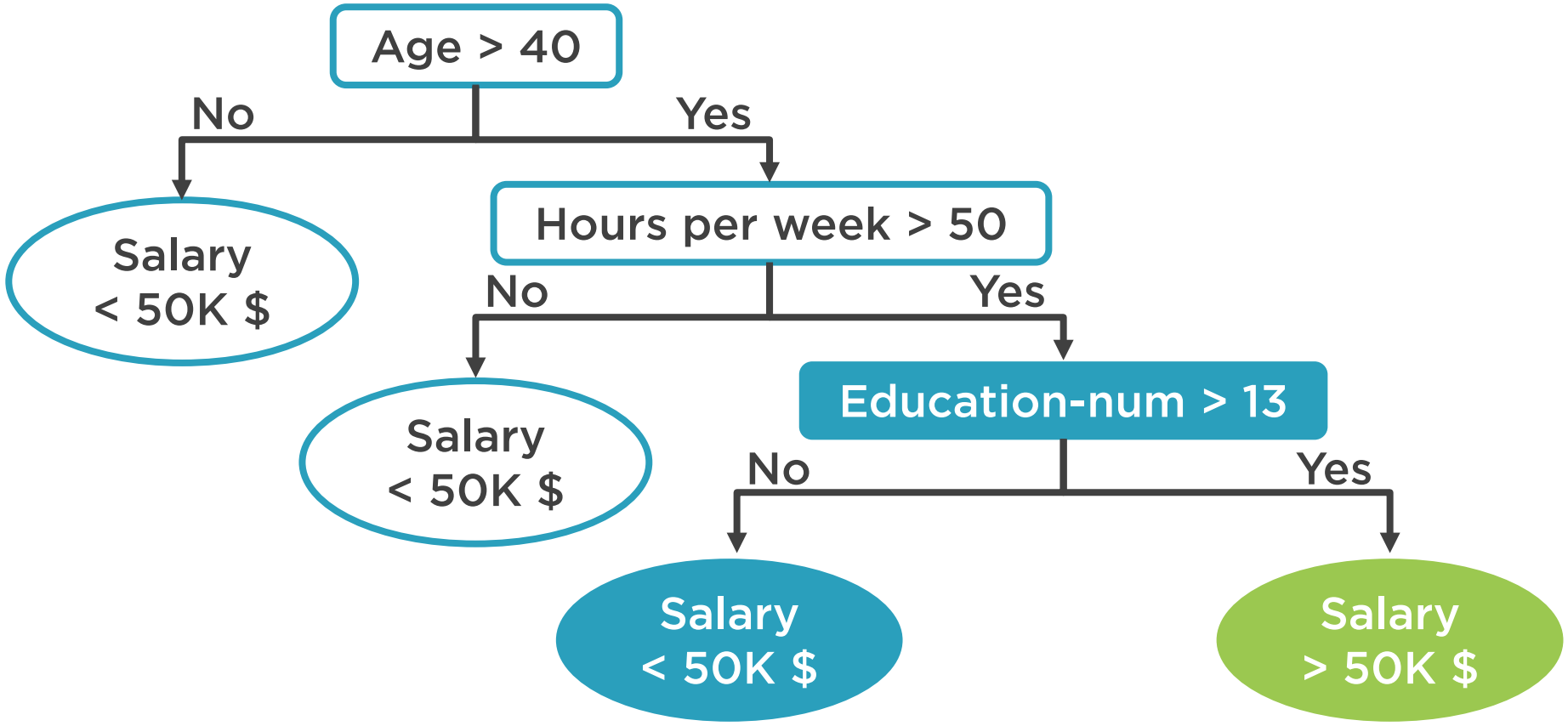
What if Data Is Missing at Prediction Time?



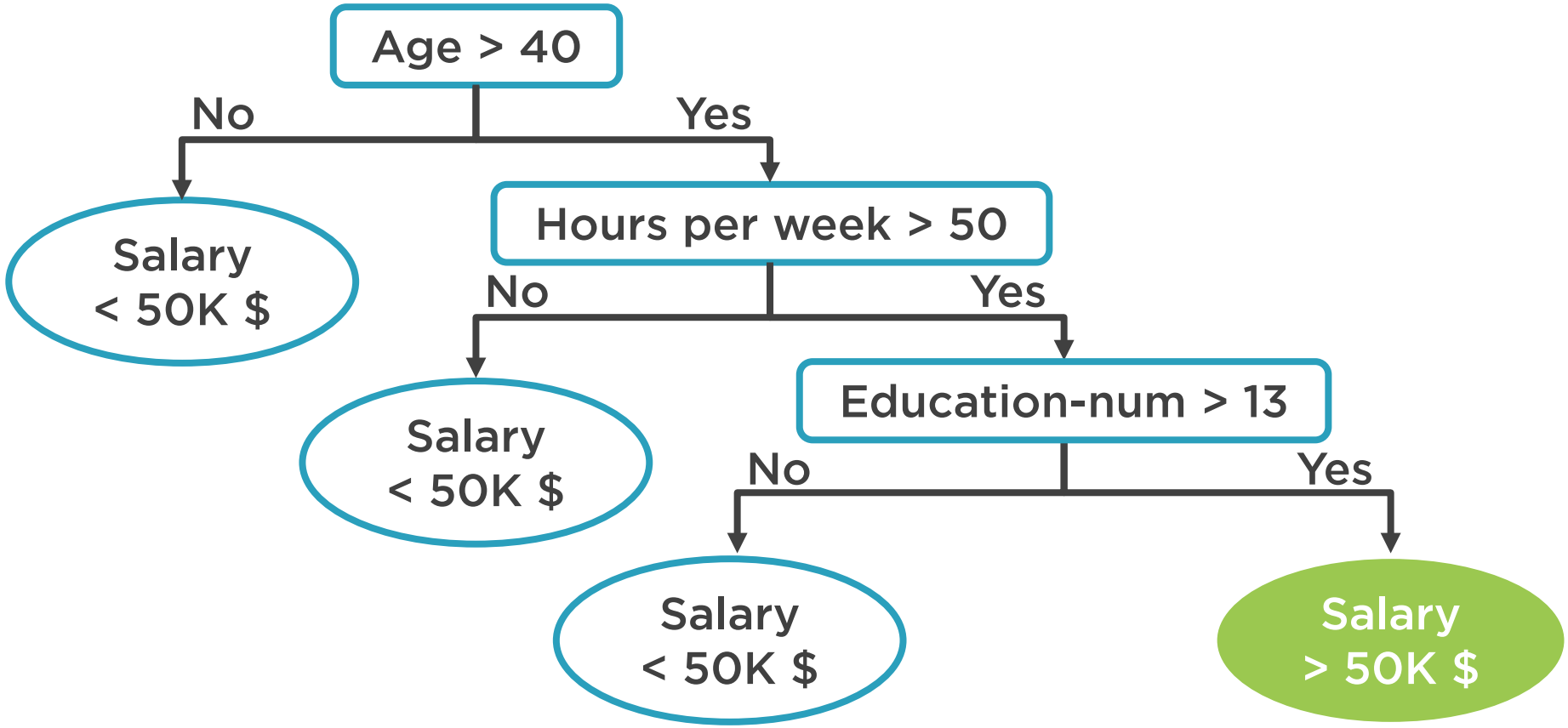


What if Data Is Missing at Prediction Time?

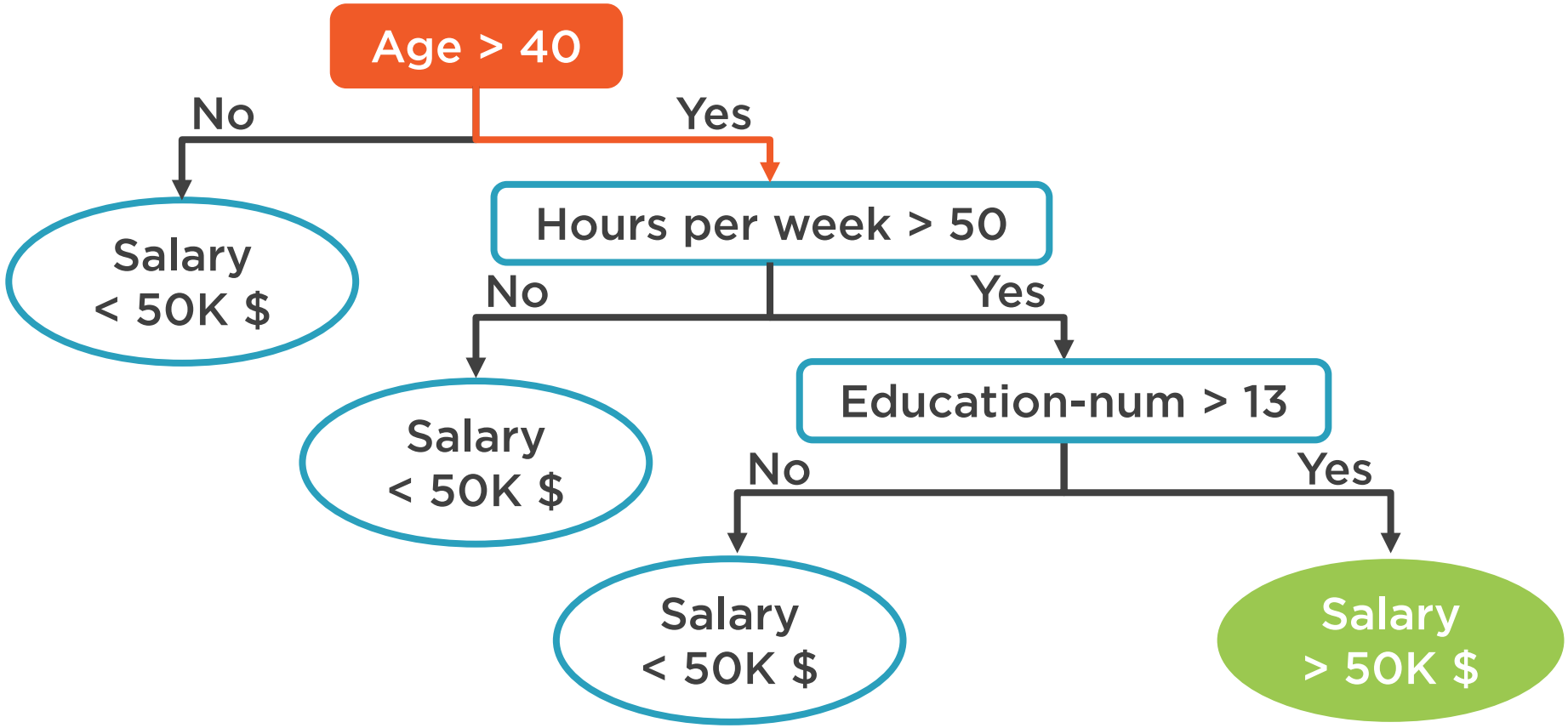




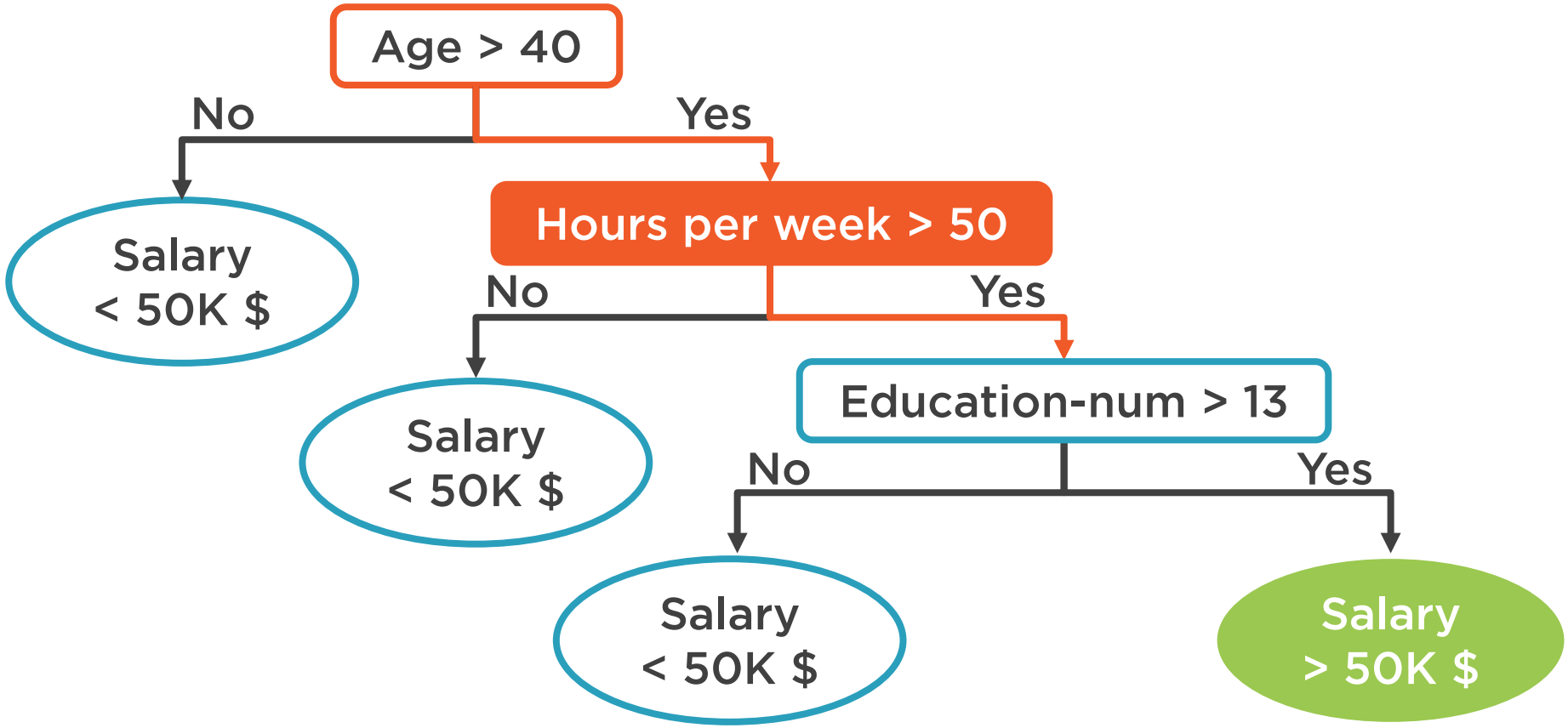
What if Data Is Missing at Prediction Time?



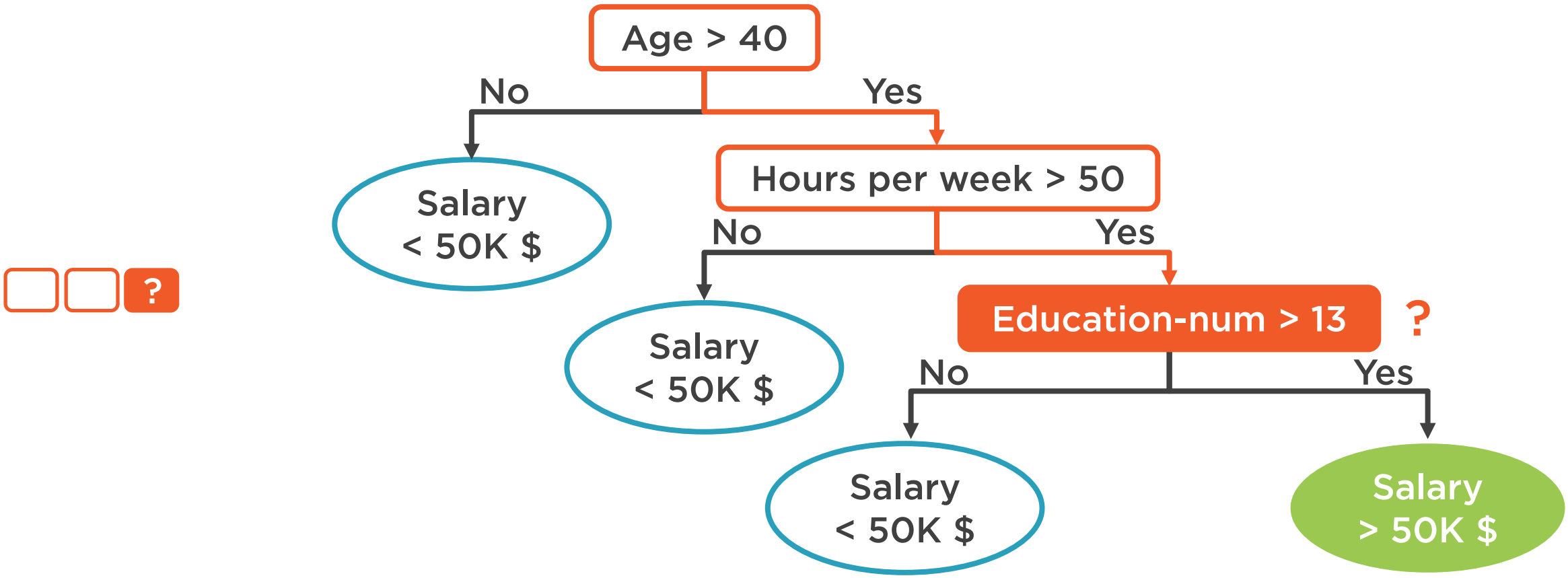
What if Data Is Missing at Prediction Time?



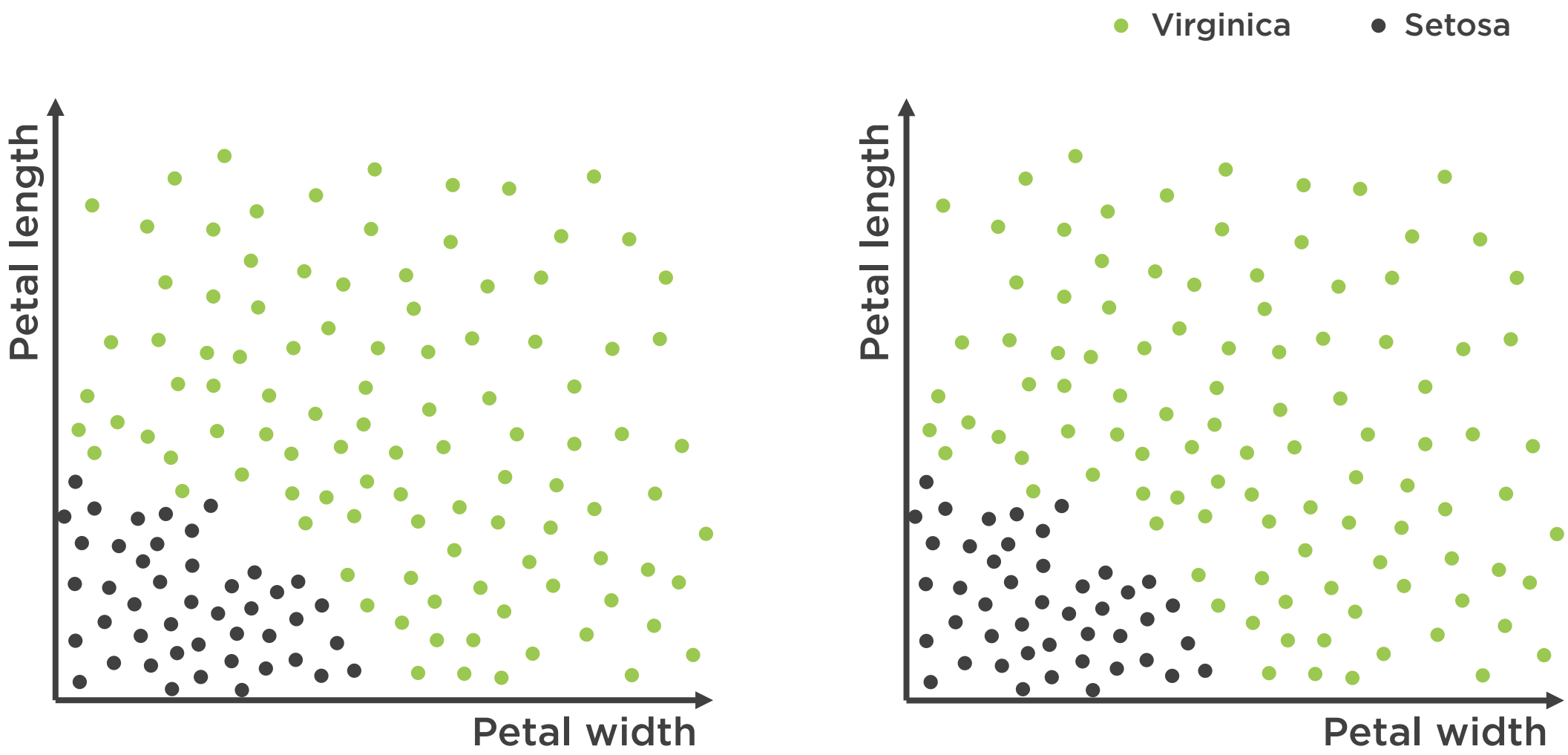
What if Data Is Missing at Prediction Time?



What if Data Is Missing at Prediction Time?

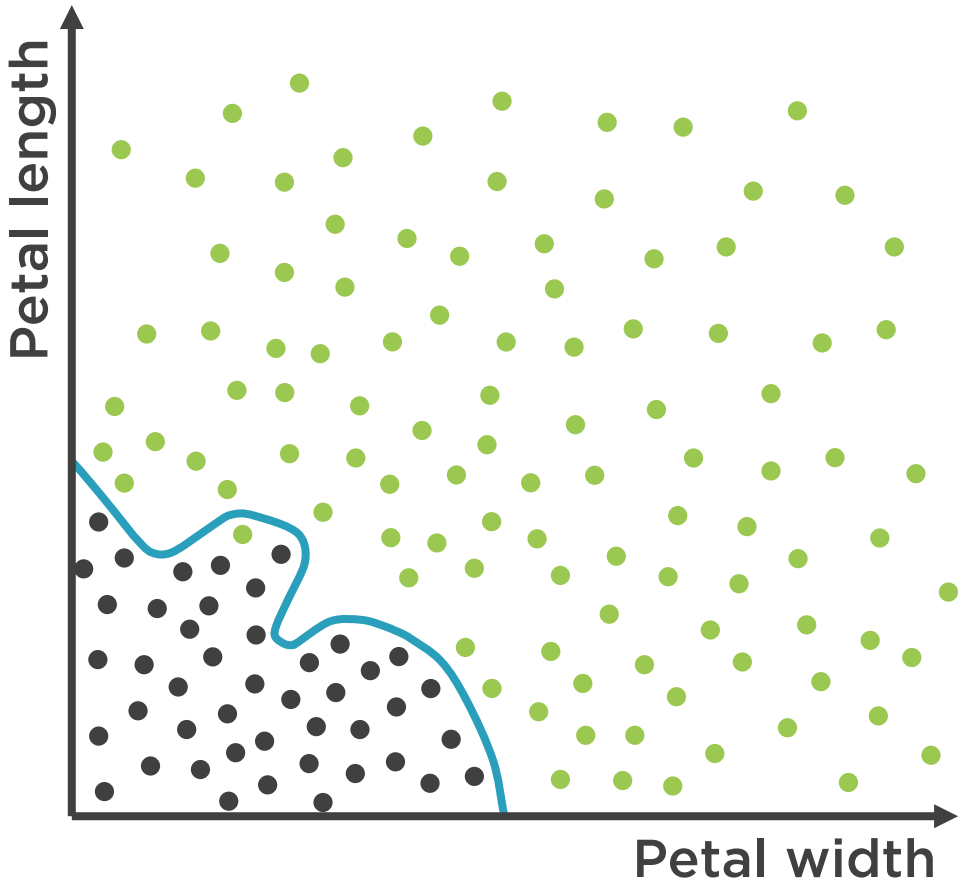


Loss of Data Impacts High Variance Models

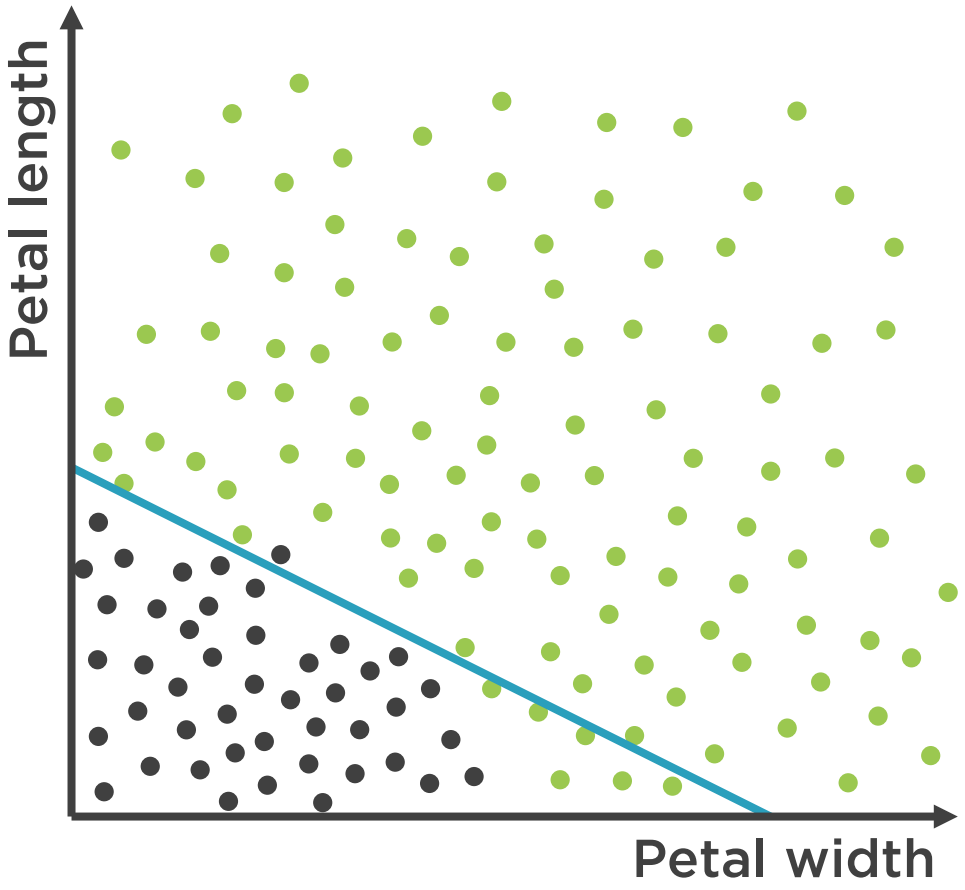


Loss of Data Impacts High Variance Models

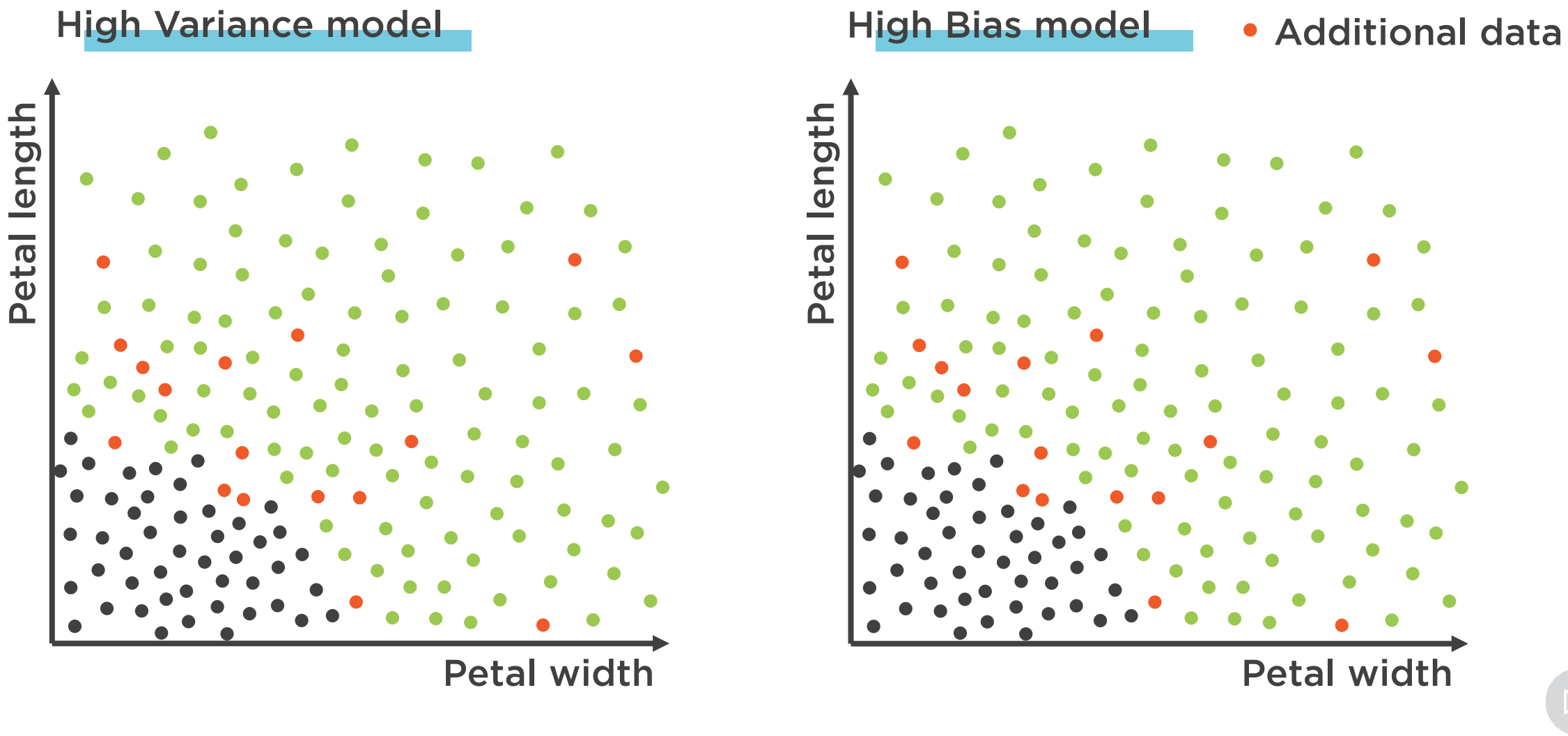
High Variance model



High Bias model

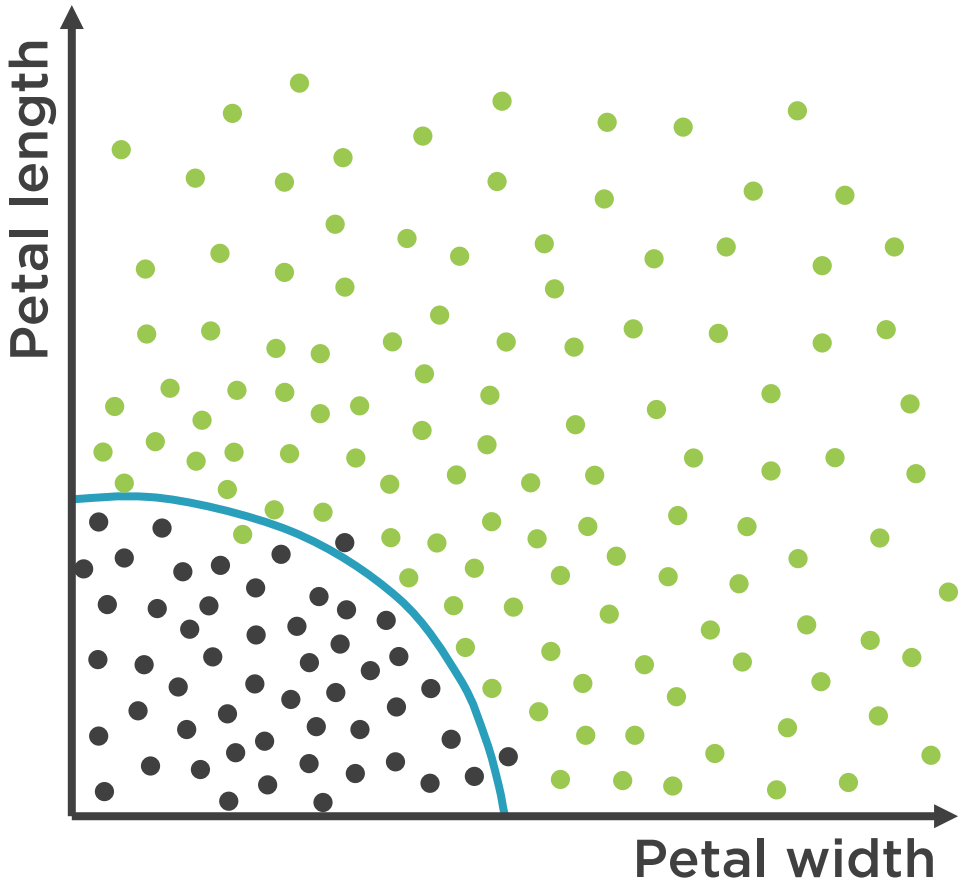


Loss of Data Impacts High Variance Models

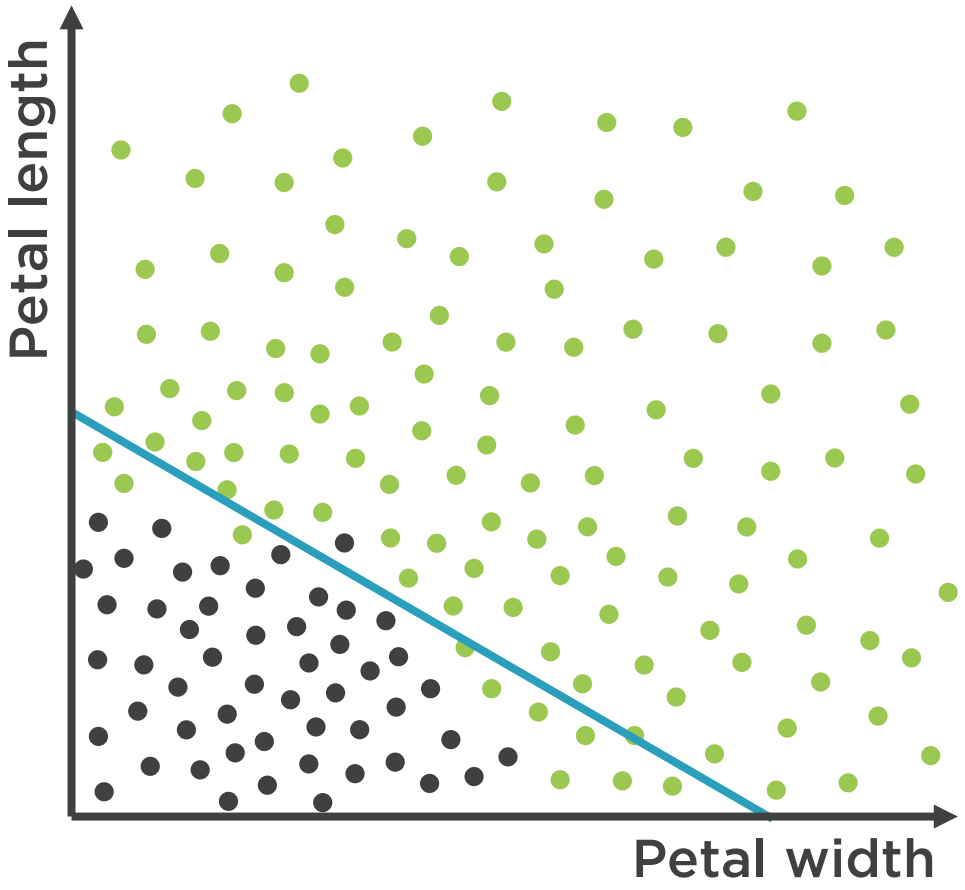


Loss of Data Impacts High Variance Models

High Variance model

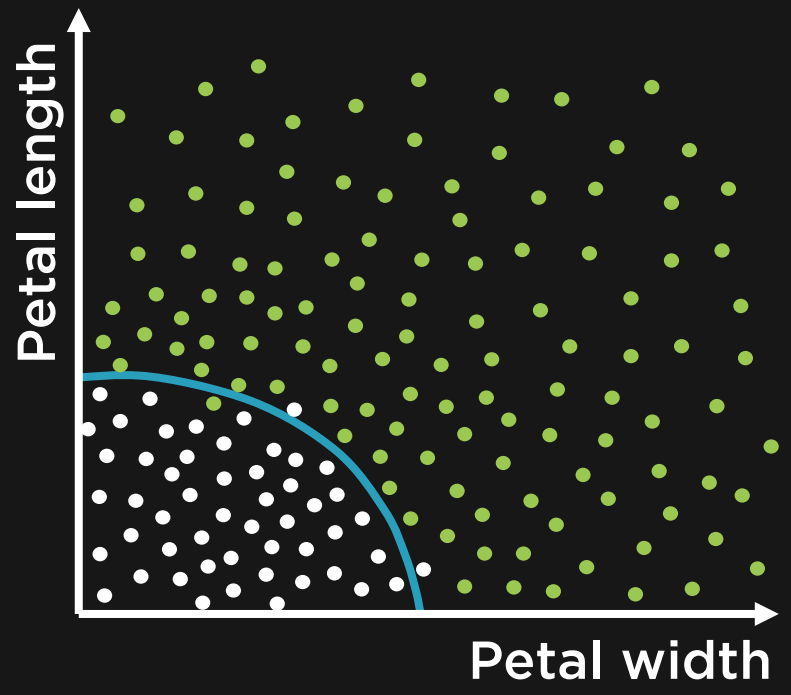


High Bias model



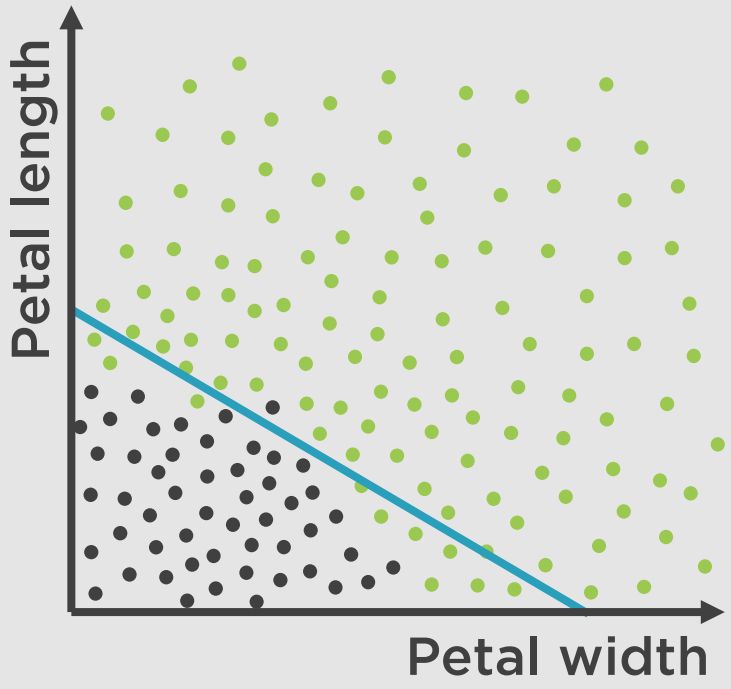
Loss of Data Impacts High Variance Models

High Variance model



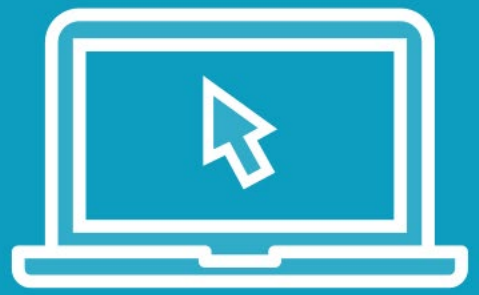
Additional Observations help the model

High Bias model



Additional Features help the model

Demo



Using Indicator variables

Replace with Mean, Median and Mode

Levels of Measurement

Sunny

Thunder

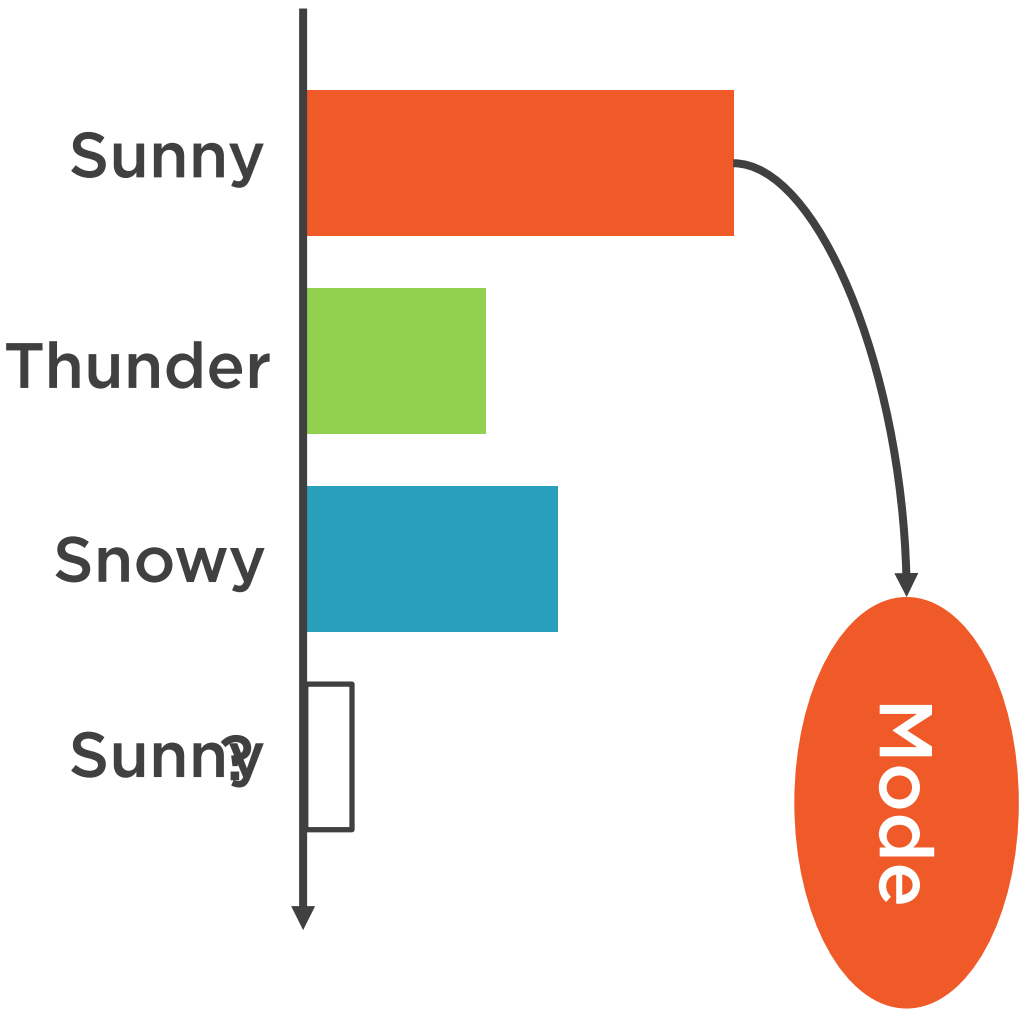
Snowy

Rainy

Name

Nominal

Name



Levels of Measurement

Nominal

Levels of Measurement

Name + Order

Nominal

Ordinal

Levels of Measurement

Name + Order + Difference

Time

But, 6/2 ≠ 16/12

Nominal

Ordinal

Interval

Levels of Measurement

Name + Order + Difference + Zero

Heights of people

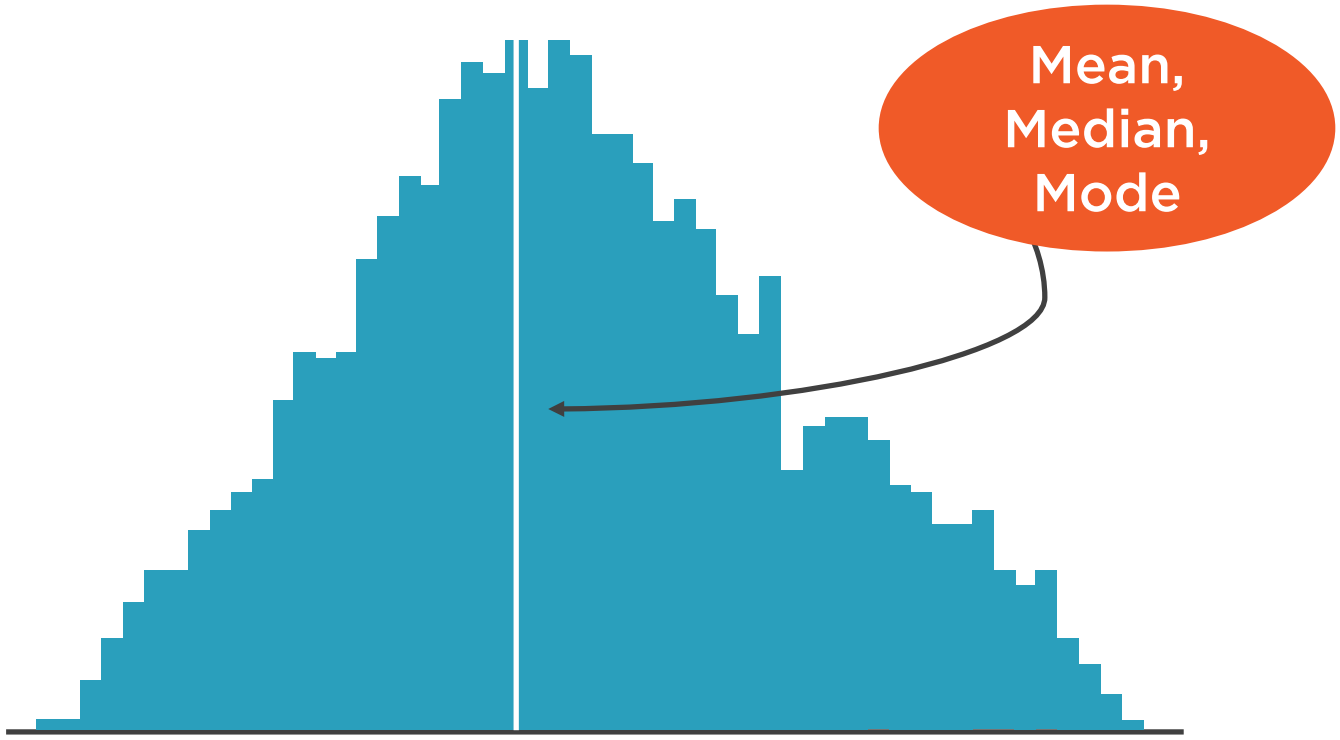
Nominal

Ordinal

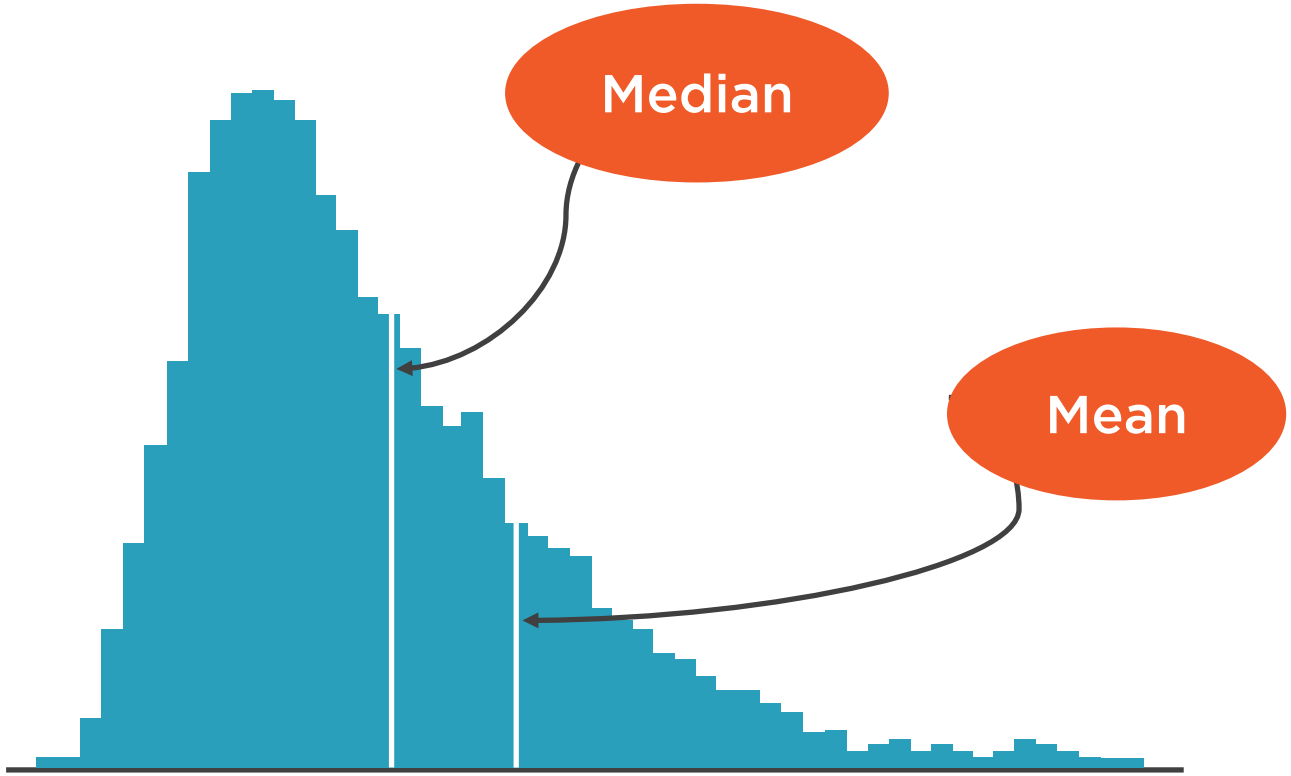
Interval

Ratio

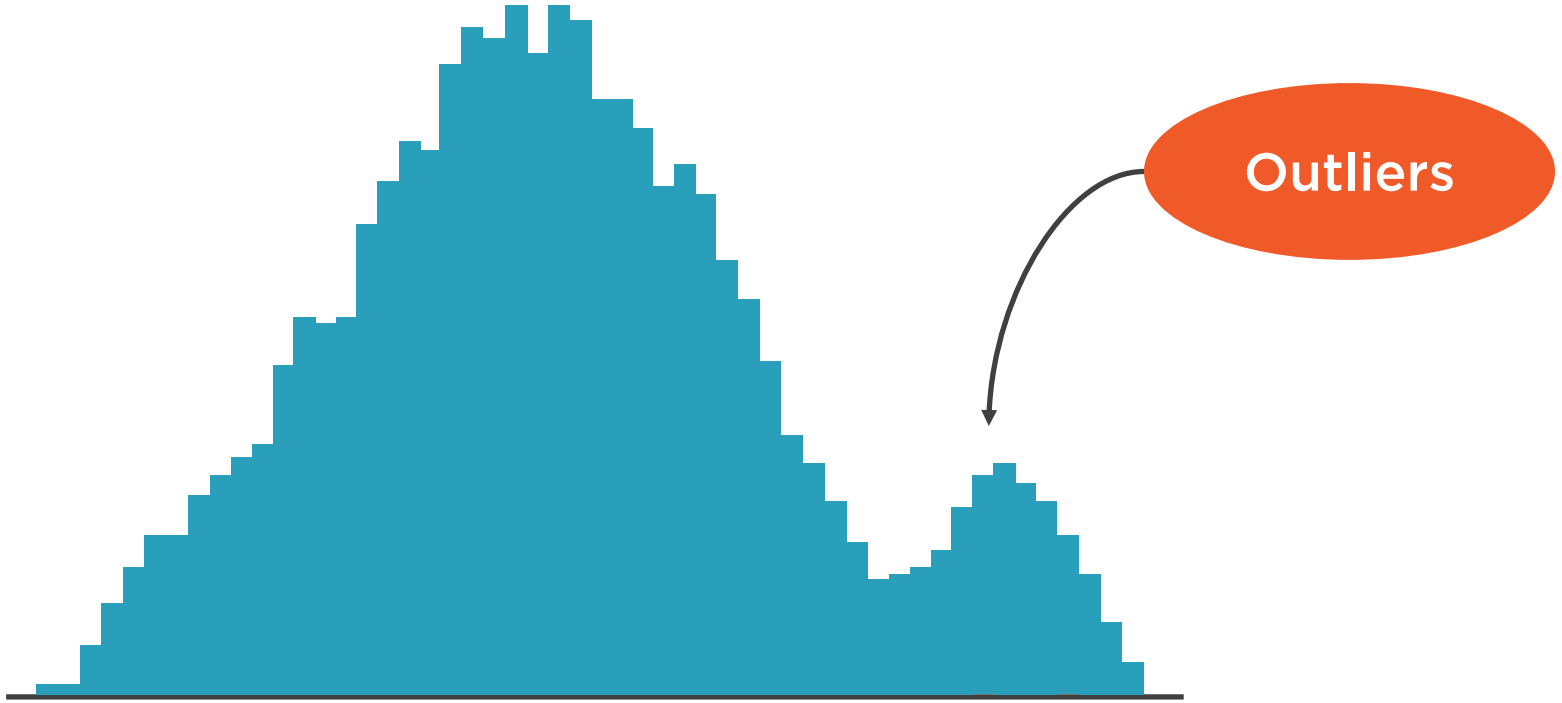
Mean or Median?



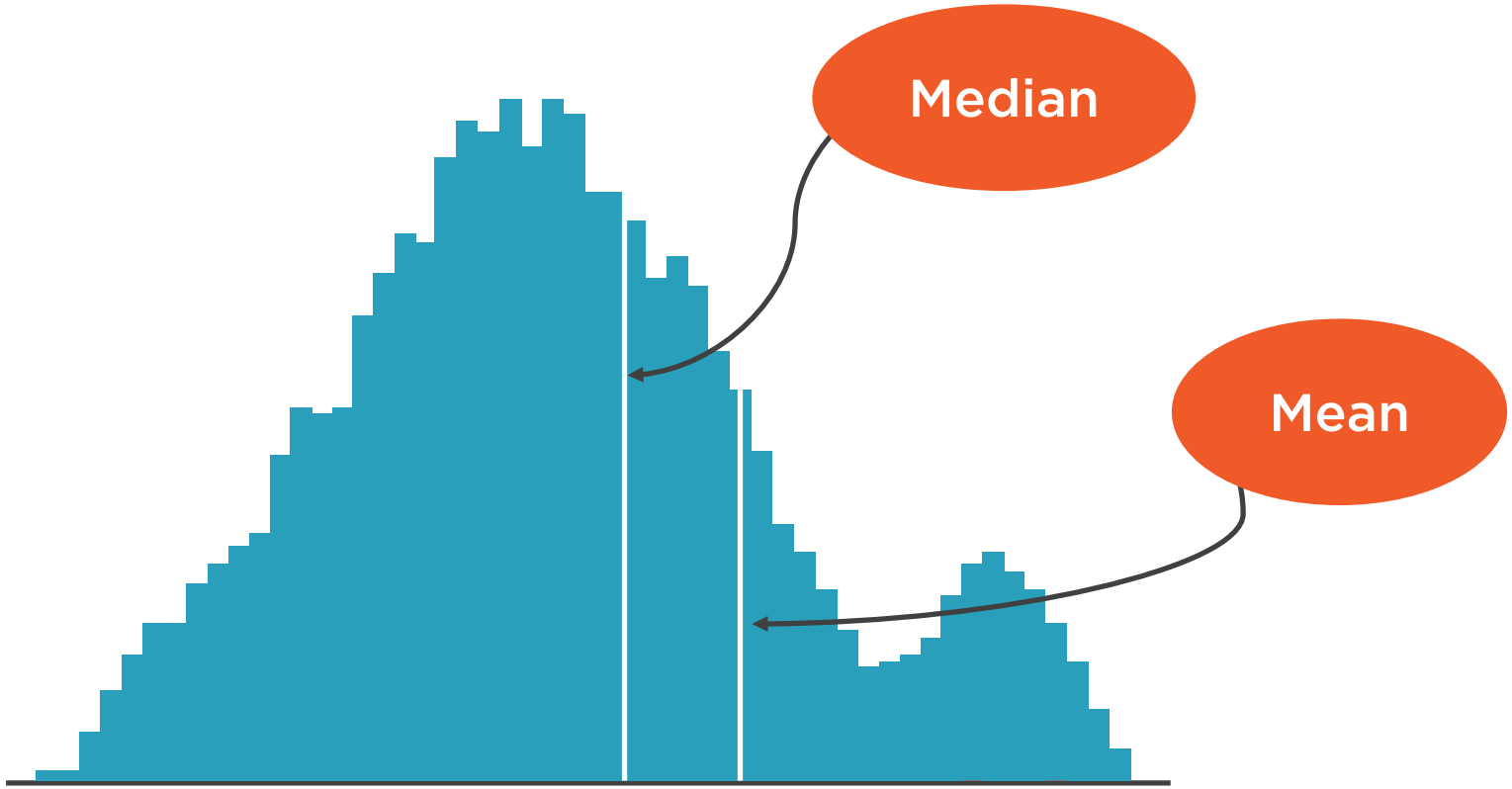
Mean or Median?



Mean or Median?



Mean or Median?



Disadvantages of Single Imputation Methods

Single Imputation methods do not preserve relationships between variables

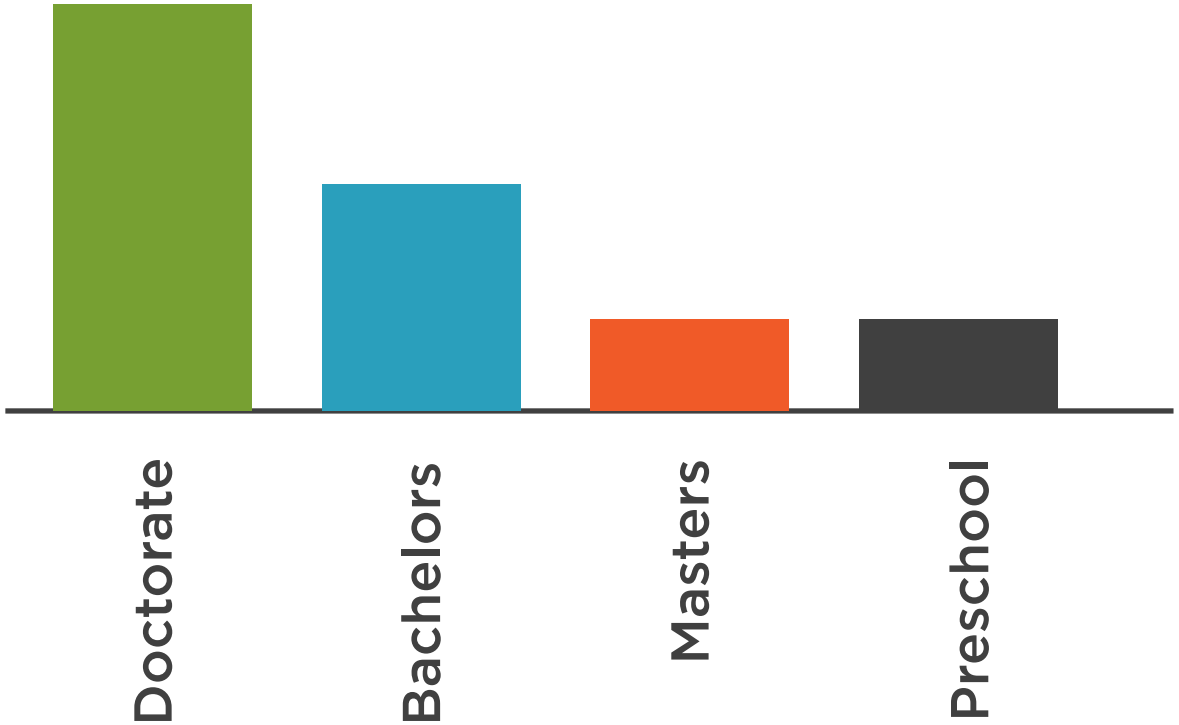
Disadvantages of Single Imputation Methods

Education	Label
Doctorate	1
1st-4th	0
Masters	1
Bachelors	1
Doctorate	1
Bachelors	0
10th	1
Preschool	0
Doctorate	1
Preschool	0

Disadvantages of Single Imputation Methods

Education	Label
Doctorate	1
?	0
Masters	1
Bachelors	1
Doctorate	1
Bachelors	0
?	1
Preschool	0
Doctorate	1
?	0

Disadvantages of Single Imputation Methods

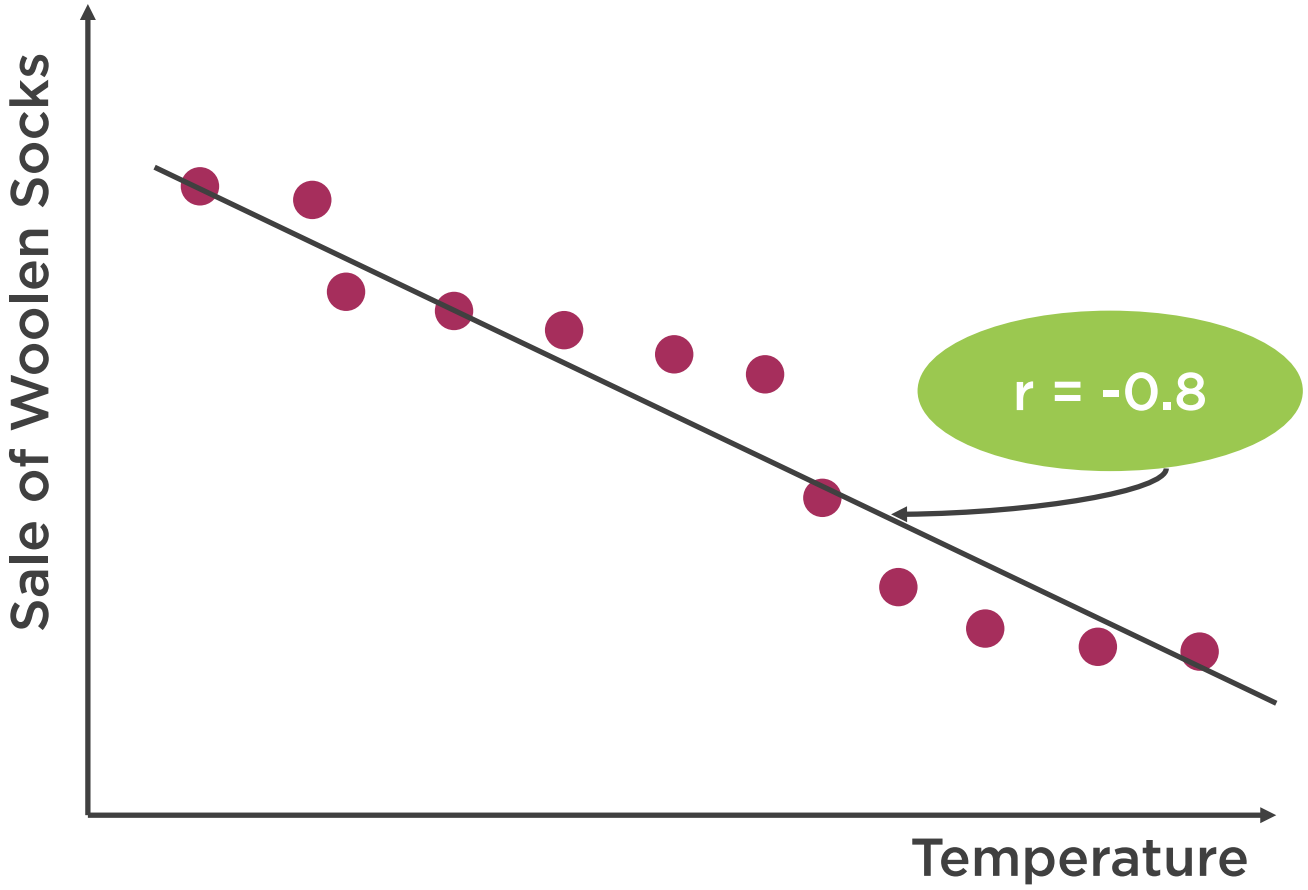


Disadvantages of Single Imputation Methods

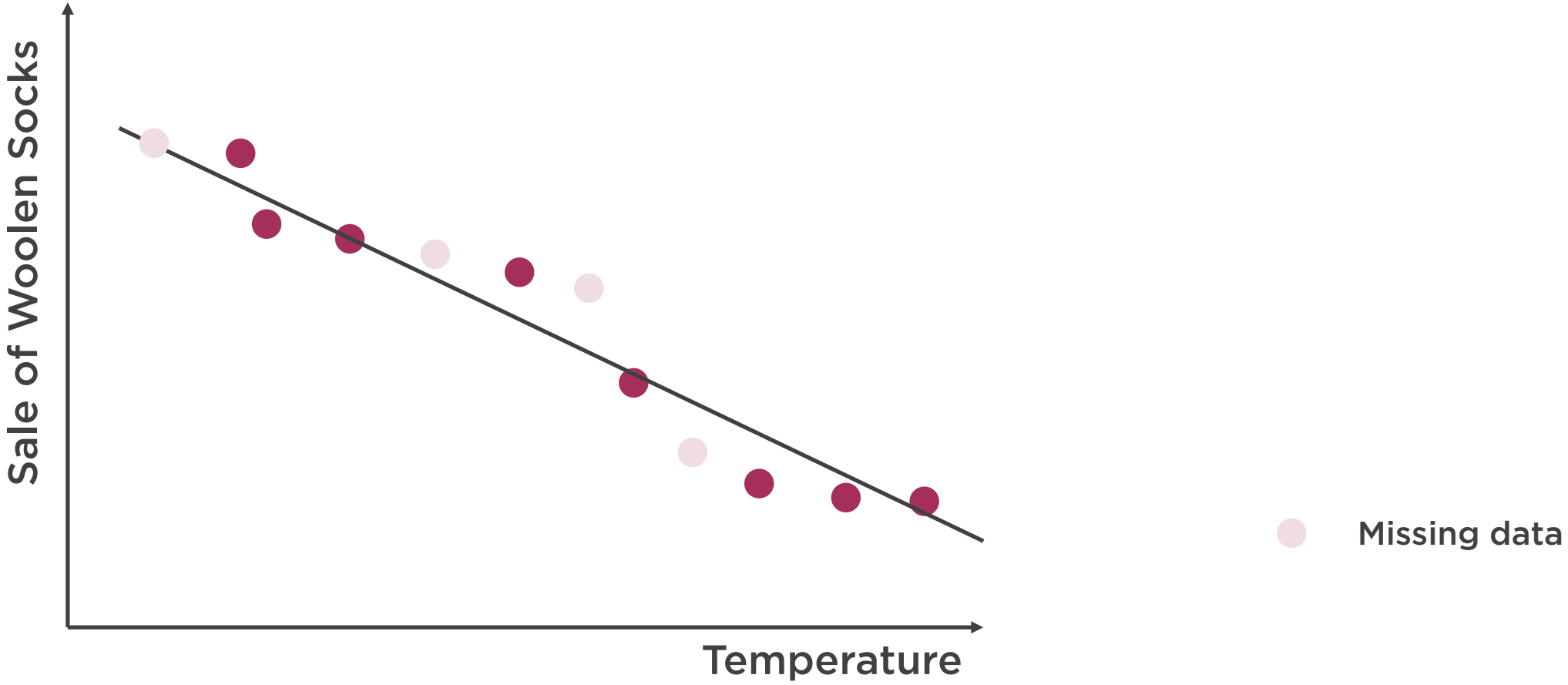
Education	Label
Doctorate	1
Doctorate	0
Masters	1
Bachelors	1
Doctorate	1
Bachelors	0
Doctorate	1
Preschool	0
Doctorate	1
Doctorate	0

So, it weakens the relationship with other variables

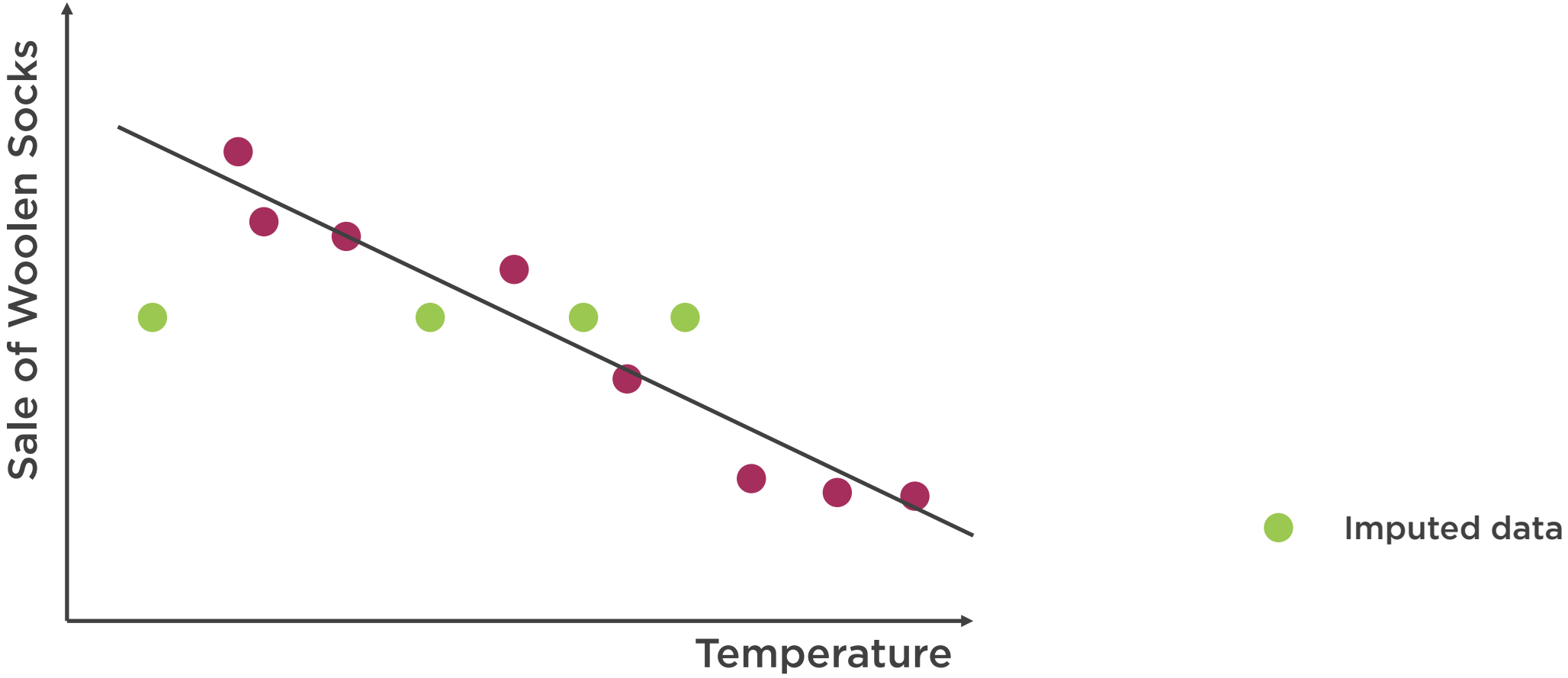
Disadvantages of Single Imputation Methods



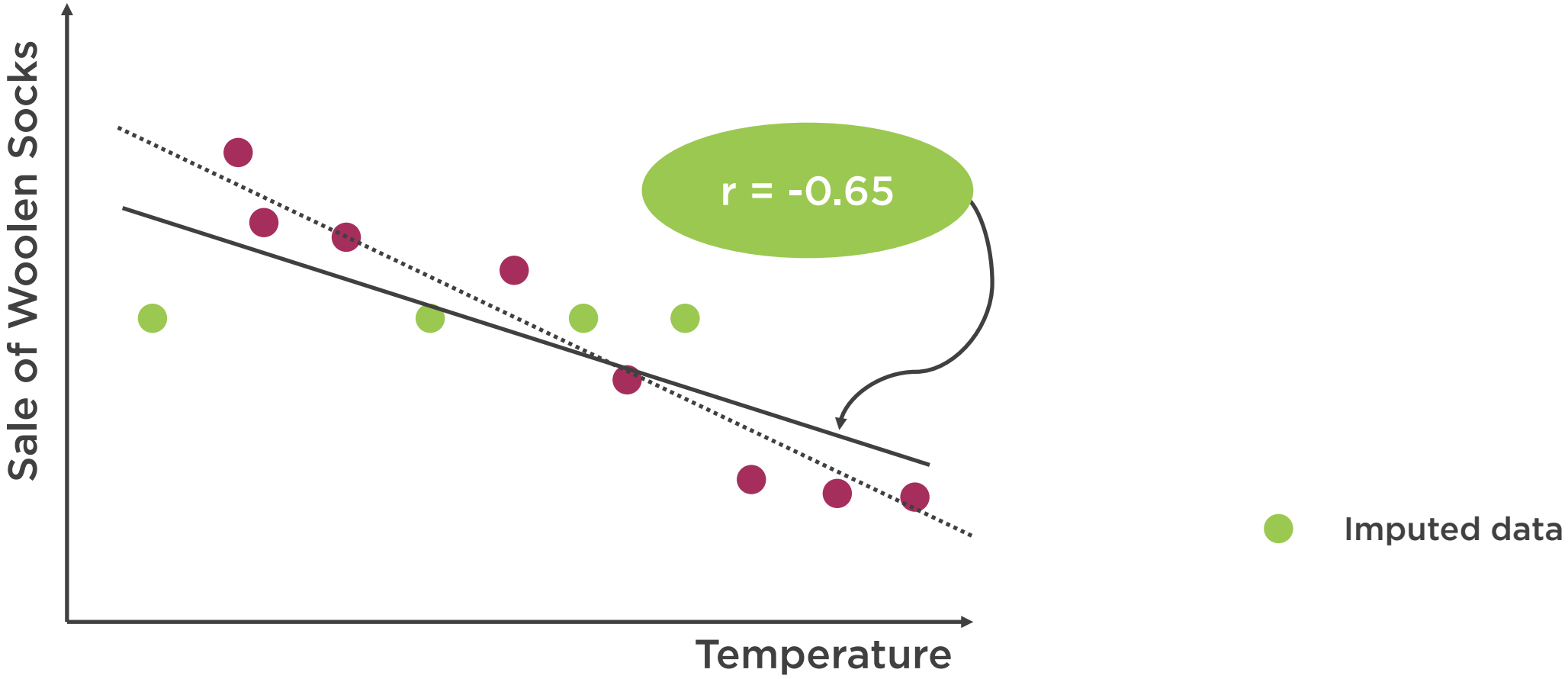
Disadvantages of Single Imputation Methods



Disadvantages of Single Imputation Methods

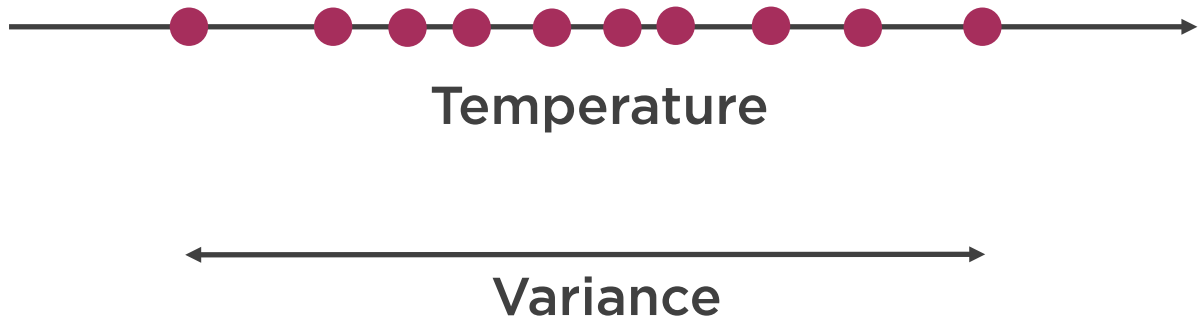


Disadvantages of Single Imputation Methods

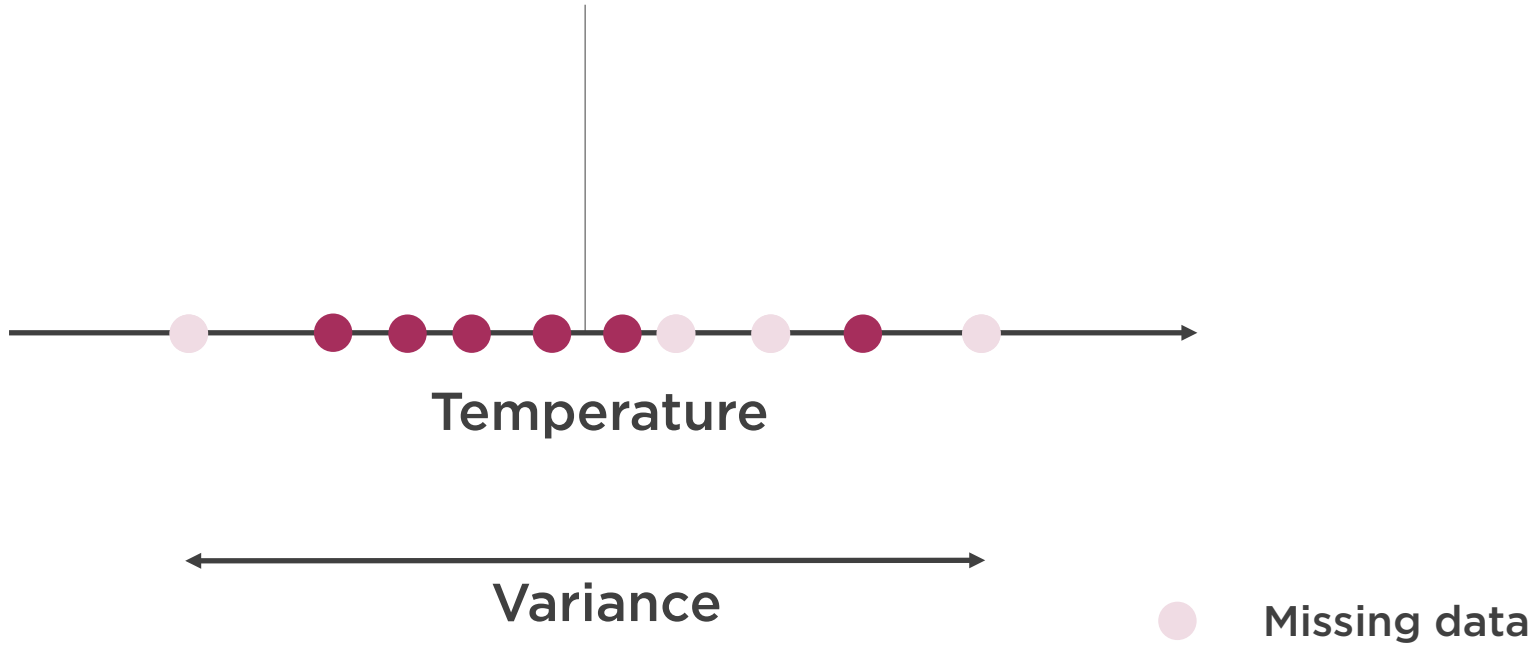


Reduces variance in the dataset

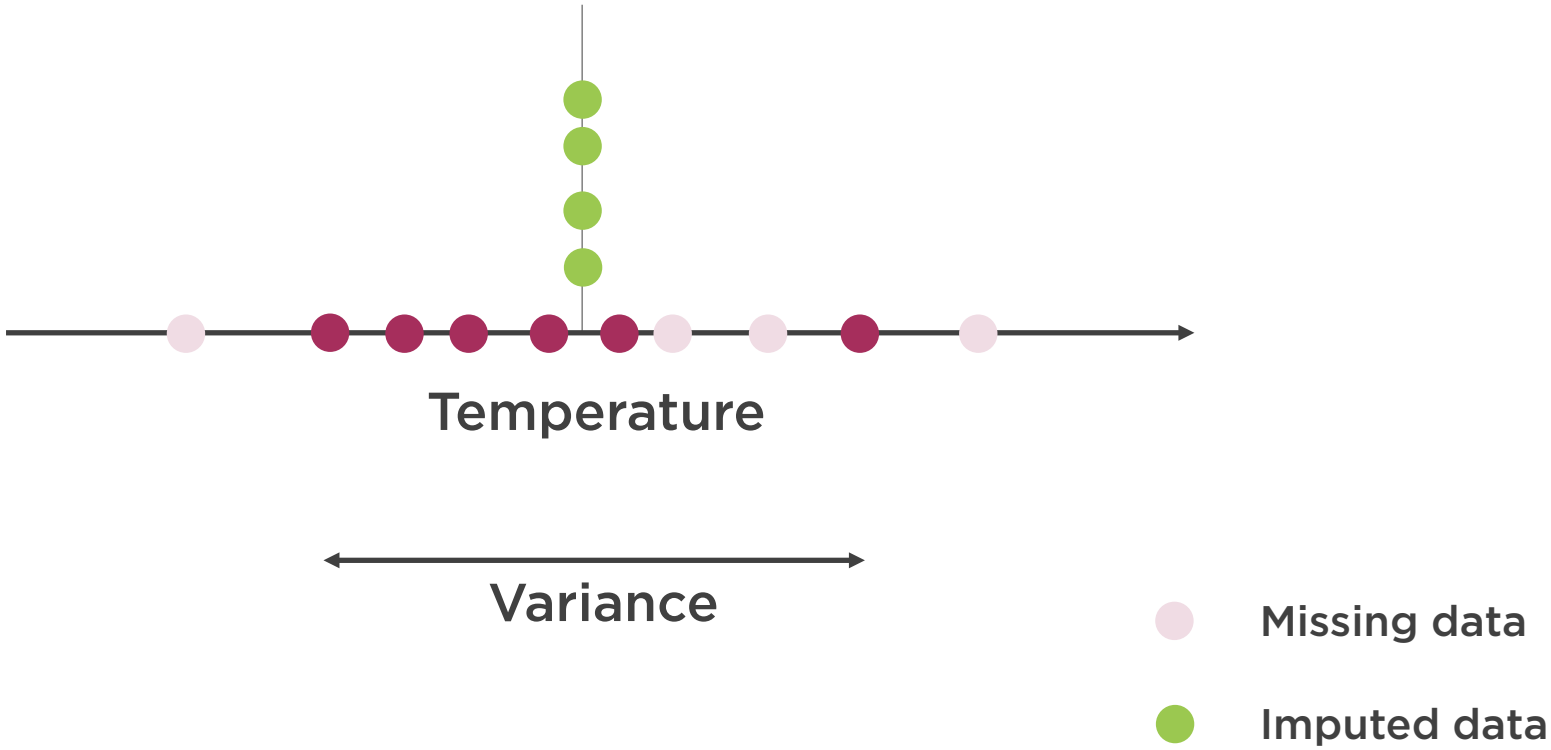
Disadvantages of Single Imputation Methods



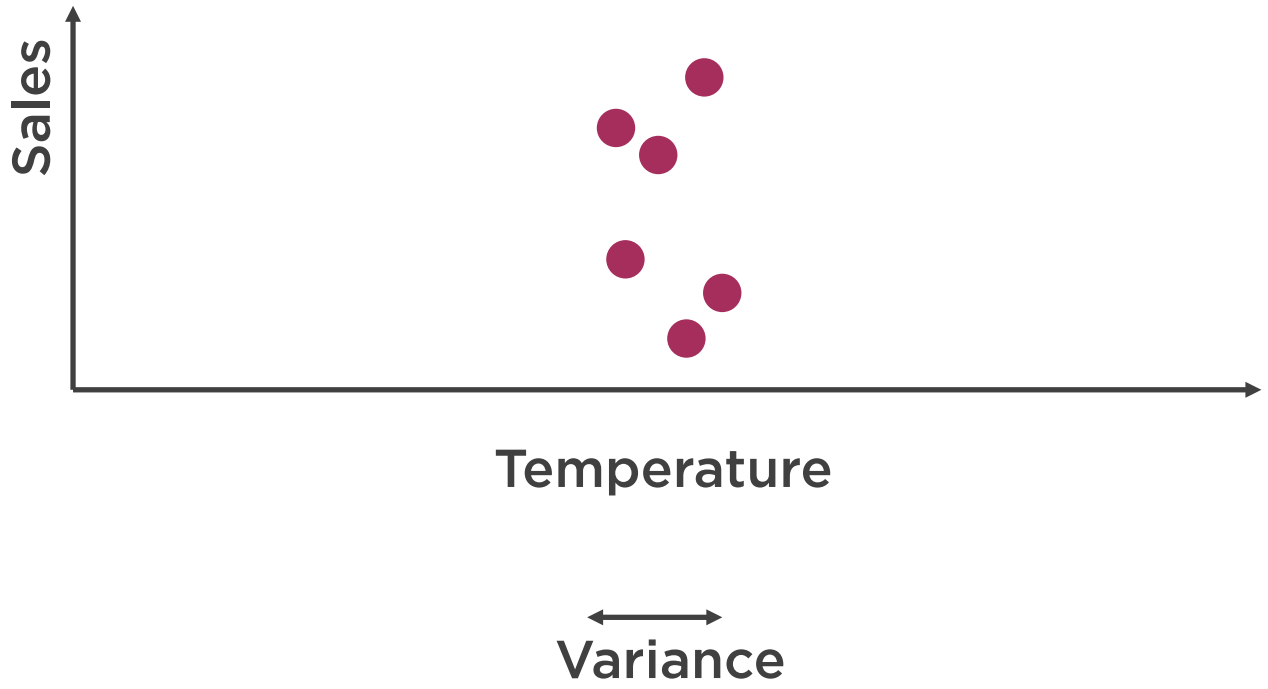
Disadvantages of Single Imputation Methods



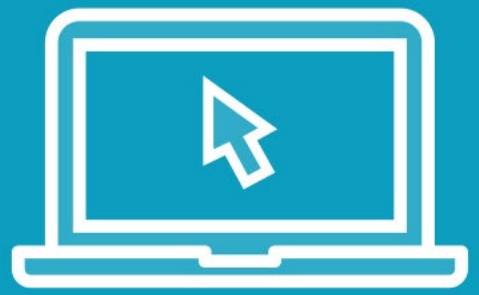
Disadvantages of Single Imputation Methods



Disadvantages of Single Imputation Methods



Demo



Multiple Imputation By Chained Equations

How MICE Works?

Multiple Imputation by Chained Equations

Age	Income	Gender
33	?	M
18	$40-60K	M
15	$60-80K	F
?	$40-60K	F

Single Imputation

Multiple Imputation by Chained Equations

Age	Income	Gender
33	?	M
18	$40-60K	M
15	$60-80K	F
20.5	$40-60K	F

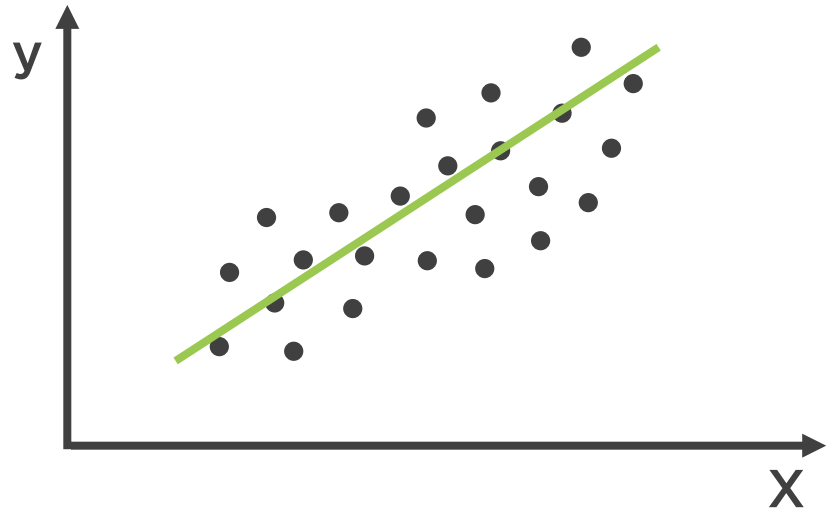
Multiple Imputation by Chained Equations

Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
20.5	$40-60K	F

Age back to ‘?’

Multiple Imputation by Chained Equations

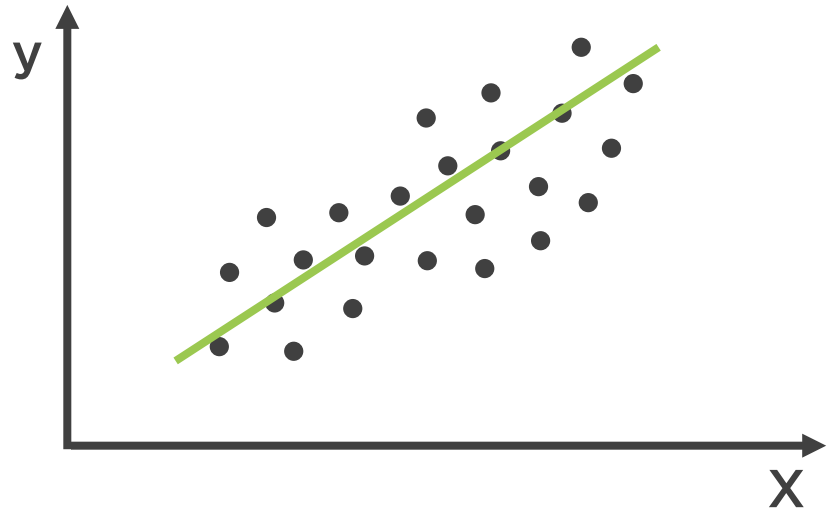
Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
?	$40-60K	F



Age ~ f(Income, Gender)

Multiple Imputation by Chained Equations

Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
35.3	$40-60K	F

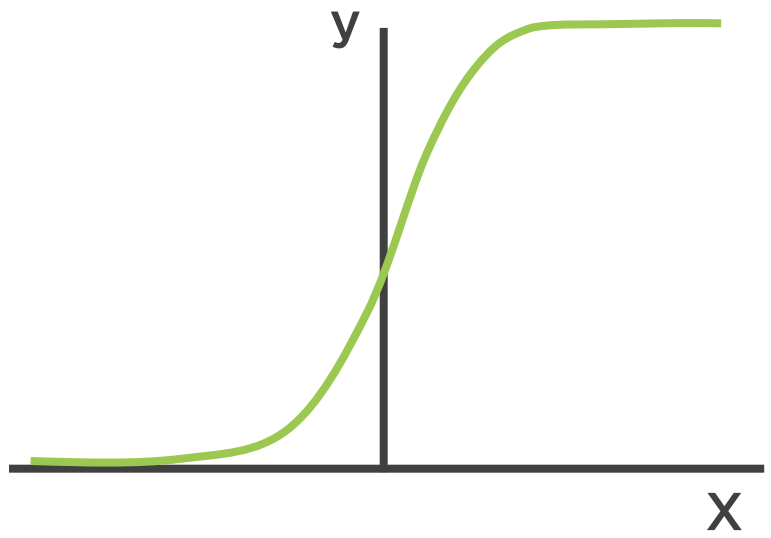


Age ~ f(Income, Gender)

Income back to ‘?’

Multiple Imputation by Chained Equations

Age	Income	Gender
33	?	M
18	$40-60K	M
15	$60-80K	F
35.3	$40-60K	F

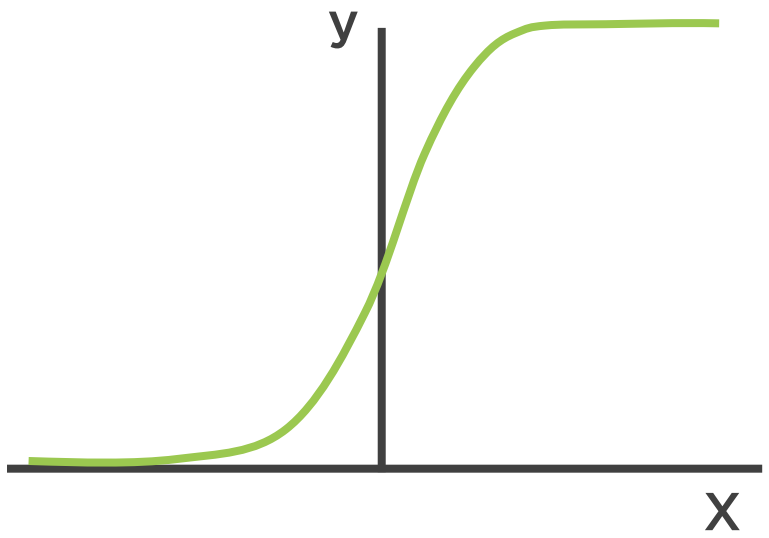


Income ~ f(Age, Gender)

Multiple Imputation by Chained Equations

Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
35.3	$40-60K	F

Income ~ f(Age, Gender)



Multiple Imputation by Chained Equations

Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
35.3	$40-60K	F

Age	Income	Gender
33	$40-60K	M
18	$40-60K	M
15	$60-80K	F
34.2	$40-60K	F

Age	Income	Gender
33	$60-80K	M
18	$40-60K	M
15	$60-80K	F
34.0	$40-60K	F

Age	Income	Gender
33	$60-80K	M
18	$40-60K	M
15	$60-80K	F
33.8	$40-60K	F

n=4

Summary

Asking questions help

Missing data methods influence accuracy

Try different methods to test assumptions