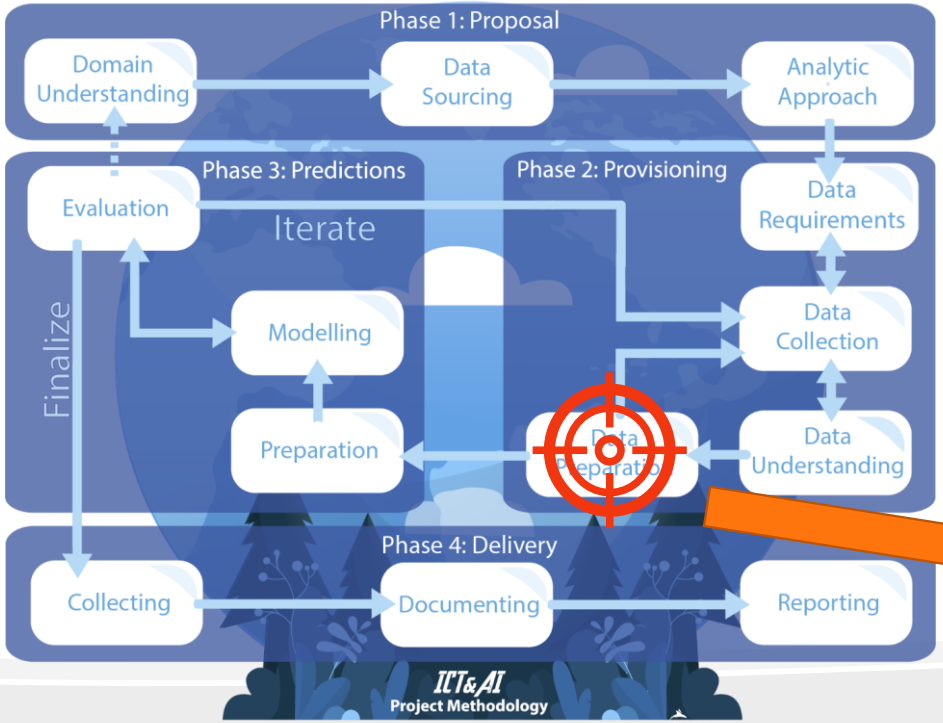

M ss ng d t

AI project methodology: your roadmap



Target variable, model and domain requirements, data source(s), ...
What do you want to achieve (predict)?

What data (quality) is required?
Define a data dictionary

How do you get (generate) and combine your data? Capture the process.

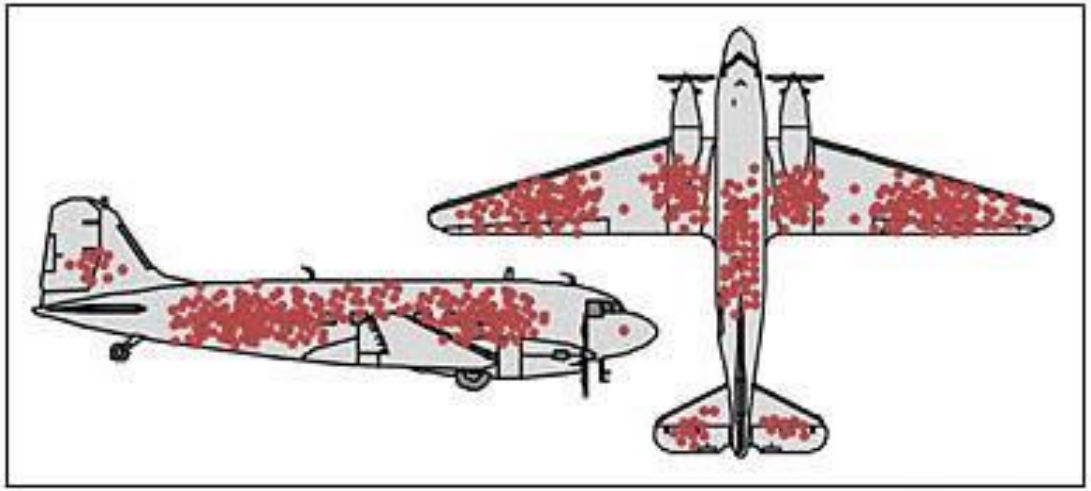
Explore your data (EDA and EDV)

Prepare your data: meet the requirements

Handling missing data

- What is missing data, why is it important to handle this issue before applying any form of data analysis?
- Different types of missing data: MCAR, MAR, MNAR
- Different techniques of dealing with missing data and how to infer which techniques could potentially solve which missing data issues

Finding missing data…



Credit: Cameron Moll

No. 1246

"TITANIC"
Launch.

To be retained for admittance to Stand.

No. 1246

Launch
OF
White Star Royal Mail Triple-Screw Steamer
"TITANIC"
At BELFAST,
Wednesday, 31st May, 1911, at 12-15 p.m.

TO BE GIVEN UP AT GATE.

Now that we have a better overview of the missing samples let's make a small experiment to discover the type of the data (MCAR,MAR,MNAR) First I will divide the set into three new sets based on travel class of the passengers

```
df_class1 = df_titanic.loc[df_titanic['Pclass'] == 1]
df_class2 = df_titanic.loc[df_titanic['Pclass'] == 2]
df_class3 = df_titanic.loc[df_titanic['Pclass'] == 3]
```

```
In [9]: df_class1.isnull().sum()
Out[9]: PassengerId     0
        Survived        0
        Pclass          0
        Name            0
        Sex             0
        Age            30
        SibSp           0
        Parch           0
        Ticket          0
        Fare            0
        Cabin          40
        Embarked        2
        dtype: int64
```

```
In [10]: df_class2.isnull().sum()
Out[10]: PassengerId     0
         Survived        0
         Pclass          0
         Name            0
         Sex             0
         Age            11
         SibSp           0
         Parch           0
         Ticket          0
         Fare            0
         Cabin         168
         Embarked        0
         dtype: int64
```

```
In [11]: df_class3.isnull().sum()
Out[11]: PassengerId     0
         Survived        0
         Pclass          0
         Name            0
         Sex             0
         Age           136
         SibSp           0
         Parch           0
         Ticket          0
         Fare            0
         Cabin         479
         Embarked        0
         dtype: int64
```

We also have 2 missing values which are from the column "Embarked" and in my opinion this is a MCAR but we don't have missing values form this column in the other classes,so there is the possibility that this event is again connected with the class separation on the ship.

Want we just discover is that the missing data in column "Age" and "Cabin" is MAR depending to the class of the passengers and as we expected we observe the most missing value in the 3rd class.

4. Embarked

The embarked column shows the passengers' city of embarkation. In this column most of the cells are filled in but there are still two cells without values. I believe that the type of the mechanism is **MNAR**. Since, I could not group it based on MCAR or MAR. The two survived passengers, Ms. Amelie Icard and Mrs. George Nelson, who have these empty cells are both in the first class and in the same cabin, B28. I am assuming that they were rich and did not want their city of embarkation stored in the database. Therefore, I believe the best way to approach this missing data is **imputation**. We can fill the data easily by asking the survived passengers directly. On the other hand, I also have done some research, and apparently Ms. Amelie Icard is the maid of Mrs. George Nelson. They both boarded the Titanic on 10th of April 1912 and their city of embarkation is Southampton.

```
import pylab as pyl
titanic['Age'].hist()
pyl.show()
```

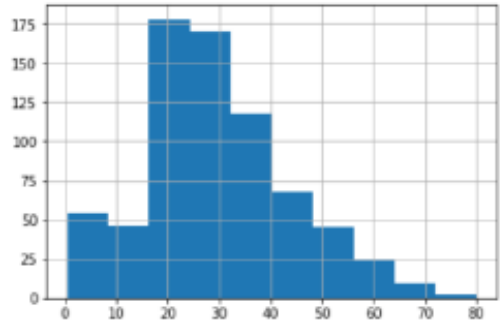


FIGURE 5: HISTOGRAM WITH MISSING VALUE

```
titanic['Age'].fillna(titanic['Age'].mean()).hist()
pyl.show()
```

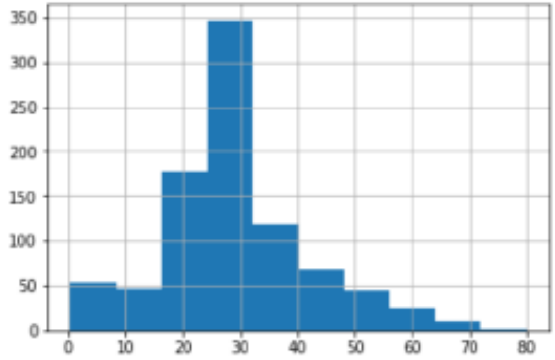


FIGURE 7: HISTOGRAM THAT IS FILLED IN WITH AVERAGE AGE

Importance of handling missing data

A. Having missing data can introduce bias in data analysis...

B. …and handling missing data incorrectly can also introduce bias…

C. …and therefore can have significant effect on the conclusions that can be drawn from the data

D. Missing data has impact on feature engineering and the use of machine learning models

E. … anything missing? ☺

MCAR: Missing complete at random

The fact that a value is missing has nothing to do with the observation being studied. Examples:

A. A questionnaire was lost in the mail

B. The battery power of a sensor died

C. A blood sample might have been damaged in the lab

MAR: Missing at random

The missing variable of an observation is related to some other observed data in the model but not to the value of missing variable itself. Examples:

A. Unemployed persons might refuse to provide information on their income in a survey on dating sites

B. A dataset for medical research contains much more data on female subjects than male subjects (including age)

C. Missing IQ values for young people (who might not have been tested yet for their IQ)

MNAR: Missing not at random

The missing variable of an observation is specifically related to what is missing. Examples:

A. A person that used drugs refused a drug test (why?)

B. A rich person did not want to reveal its salary (why?)

C. A person was too ill to attend a medical screening (why?)

Different types of missing data (MCAR/MAR/MNAR)

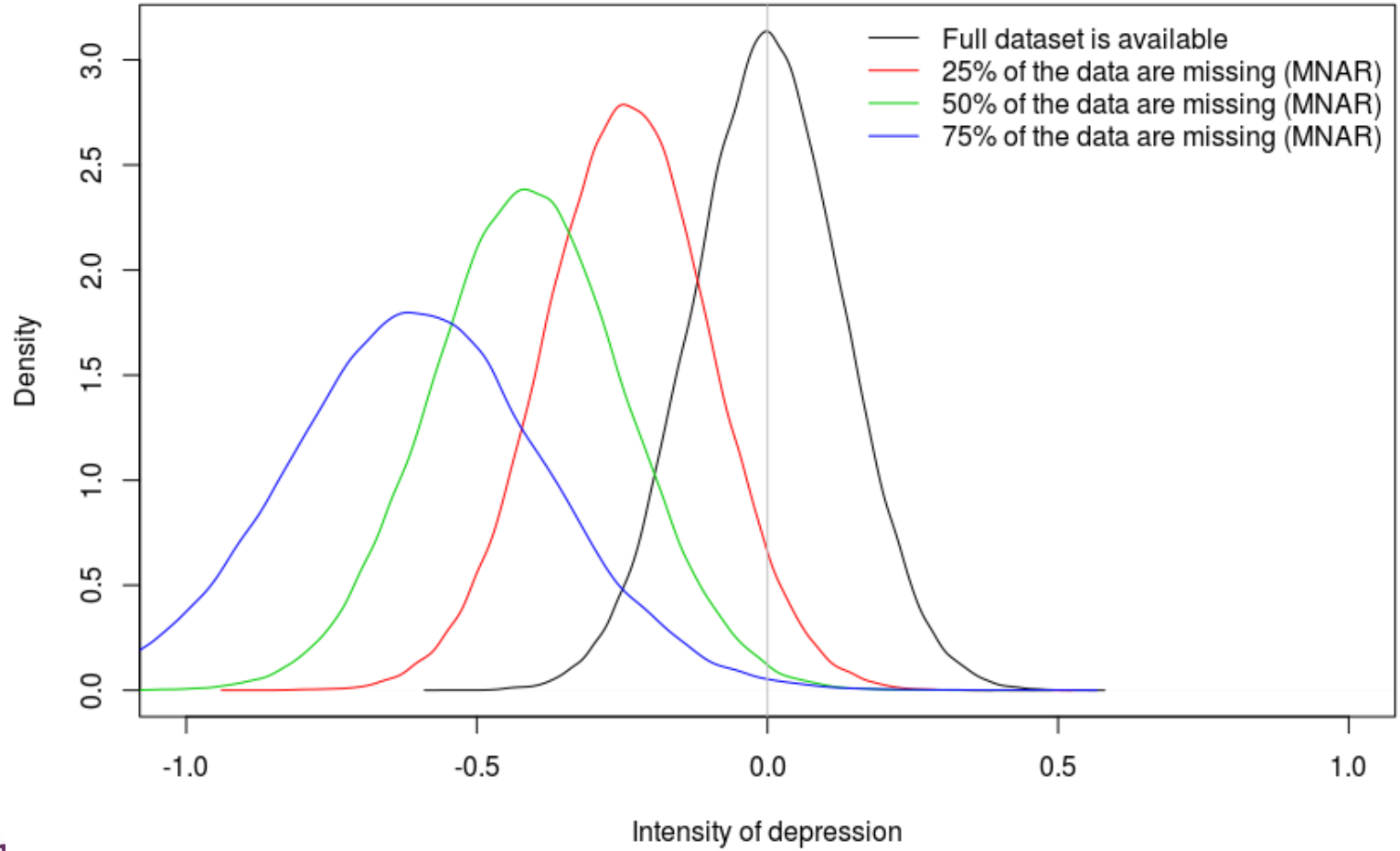
A company conducts an anonymous survey about intimate relations at work and the way they influence the working atmosphere. Many data points of one entire department appear to be missing. What is going on?

Explanation: there is some evidence that the working atmosphere in this department is under influence of complicated working relations.

Explanation: the department is notorious for mistrust in the company's policies and people tend to refuse to participate in surveys.

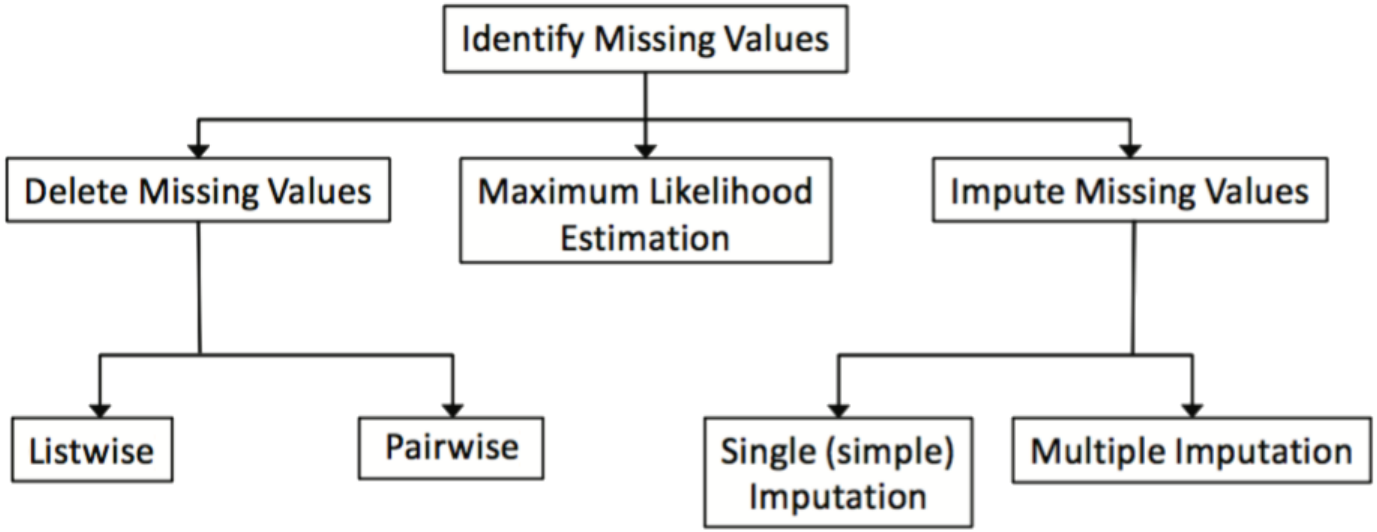
Explanation: there appears to have been a malfunction in the e-mail system and it is very unclear which employees of the department received the survey and who did not.

Explanation: the survey was conducted in an off-season period for this department when many employees were absent due to short holidays and days off.


Full dataset is available
25% of the data are missing (MNAR)
50% of the data are missing (MNAR)
75% of the data are missing (MNAR)
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-1.0
-0.5
0.0
0.5
1.0
Intensity of depression

Handling missing data…

Techniques of dealing with missing data



name	coconuts eaten	ice creams eaten	donuts eaten	weight gained
Tomas B.	17	16	18	0,3
Harry P.	8	12	11	0,5
Sintia K.	45	N/A	13	1,6
Bonnie K.	18	17	14	0,1
Harley Q.	N/A	12	15	0,9
Billy H.	12	14	16	1,1
Lizzy M.	11	N/A	39	2,6
Yukon W.	13	N/A	17	1,2
Grazia V.	42	N/A	N/A	5,6
Melodia B.	3	5	10	0,2

Listwise Deletion

name	coconuts eaten	ice creams eaten	donuts eaten	weight gained
Tomas B.	17	16	18	0,3
Harry P.	8	12	11	0,5
Sintia K.	45	N/A	13	1,6
Bonnie K.	18	17	14	0,1
Harley Q.	N/A	12	15	0,9
Billy H.	12	14	16	1,1
Lizzy M.	11	N/A	39	2,6
Yukon W.	13	N/A	17	1,2
Grazia V.	42	N/A	N/A	5,6
Melodia B.	3	5	10	0,2

Pairwise Deletion

name	coconuts eaten	ice creams eaten	donuts eaten	weight gained
Tomas B.	17	16	18	0,3
Harry P.	8	12	11	0,5
~~Sintia K.~~	~~45~~	~~N/A~~	~~13~~	~~1,6~~
Bonnie K.	18	17	14	0,1
~~Harley Q.~~	~~N/A~~	~~12~~	~~15~~	~~0,9~~
Billy H.	12	14	16	1,1
~~Lizzy M.~~	~~11~~	~~N/A~~	~~39~~	~~2,6~~
~~Yukon W.~~	~~13~~	~~N/A~~	~~17~~	~~1,2~~
~~Grazia V.~~	~~42~~	~~N/A~~	~~N/A~~	~~5,6~~
Melodia B.	3	5	10	0,2
average per person	11,6	12,8	13,8	0,44

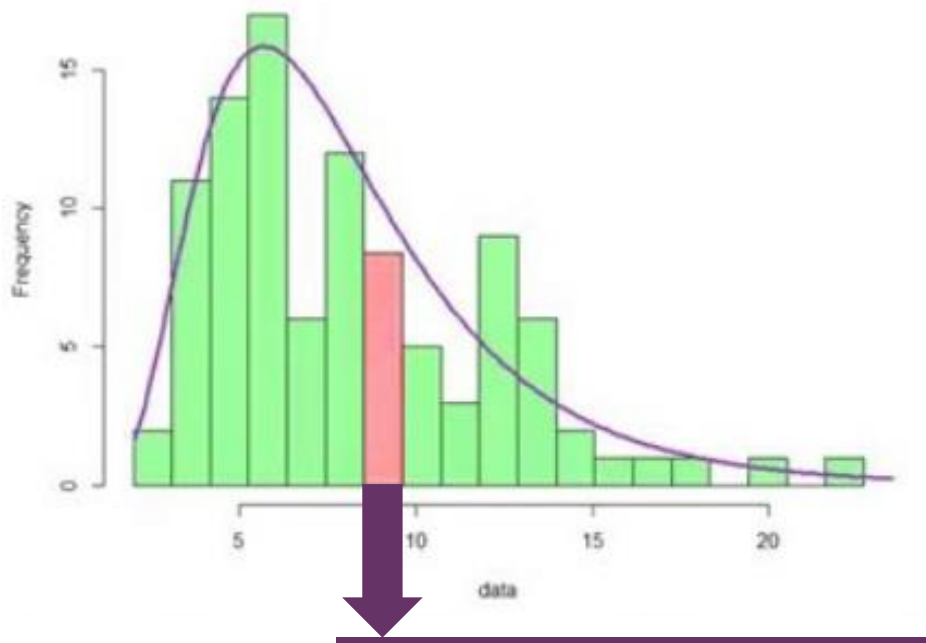
Listwise Deletion

name	coconuts eaten	ice creams eaten	donuts eaten	weight gained
Tomas B.	17	16	18	0,3
Harry P.	8	12	11	0,5
~~Sintia K.~~	~~45~~		~~13~~	~~1,6~~
Bonnie K.	18	17	14	0,1
~~Harley Q.~~		~~12~~	~~15~~	~~0,9~~
Billy H.	12	14	16	1,1
~~Lizzy M.~~	~~11~~		~~39~~	~~2,6~~
~~Yukon W.~~	~~13~~		~~17~~	~~1,2~~
~~Grazia V.~~	~~42~~			~~5,6~~
Melodia B.	3	5	10	0,2
average per person	18,8	12,7	17,0	1,41

Pairwise Deletion

The effect on the summary statistics of the data set and introduced bias depends on the chosen method of deletion

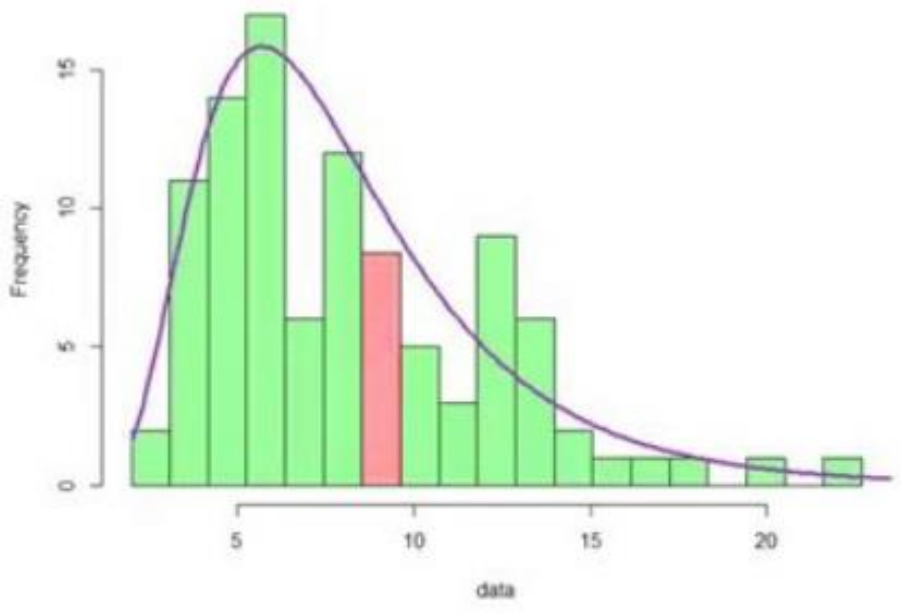
Maximum Likelihood Estimation



MONTH	SALES
January	2
February	4
March	8
April	16
May	MISSING
June	64
July	64
August	32
September	MISSING
October	MISSING
November	4
December	2

Estimate a missing value by calculating the mean and variance

Maximum Likelihood Estimation



MONTH	SALES
January	2
February	4
March	8
April	16
May	32
June	64
July	64
August	32
September	16
October	8
November	4
December	2

The estimates fit perfectly along the regression line without any residual variance. This causes relationships to be over identified.

PERSON	AGE
Robby F.	34
Nelson T.	22
Cathy B.	23
Ria Y.	26
Yin H.	
Yang C.	31
Obote W.	24
Kwame B.	25
Saladin A.	
Tito S.	
Lamia F.	23
Lady H.	22
Shania A.	33
Cyrus G.	28
Ada M.N.	29
Mona E.	21

PERSON	AGE
Robby F.	34
Nelson T.	22
Cathy B.	23
Ria Y.	26
Yin H.	34
Yang C.	31
Obote W.	24
Kwame B.	25
Saladin A.	23
Tito S.	31
Lamia F.	23
Lady H.	22
Shania A.	33
Cyrus G.	28
Ada M.N.	29
Mona E.	21

PERSON	AGE
Robby F.	34
Nelson T.	22
Cathy B.	23
Ria Y.	26
Yin H.	26
Yang C.	31
Obote W.	24
Kwame B.	25
Saladin A.	26
Tito S.	26
Lamia F.	23
Lady H.	22
Shania A.	33
Cyrus G.	28
Ada M.N.	29
Mona E.	21

Other techniques of handling missing data

A Contact original investigators to request missing data

B Find new data points by using interpolation

C Use full analysis algorithms (e.g. expectation-maximization algorithm)

… ...

Missing data take aways

CONTEXT

Use business understanding to create data context information

ANALYZE

Use context information to analyse *why* data is missing

METHODS

Know *what* methods are available and what the pros and cons are

STRATEGY

Create a plan *how* to handle the missing data and evaluate the results

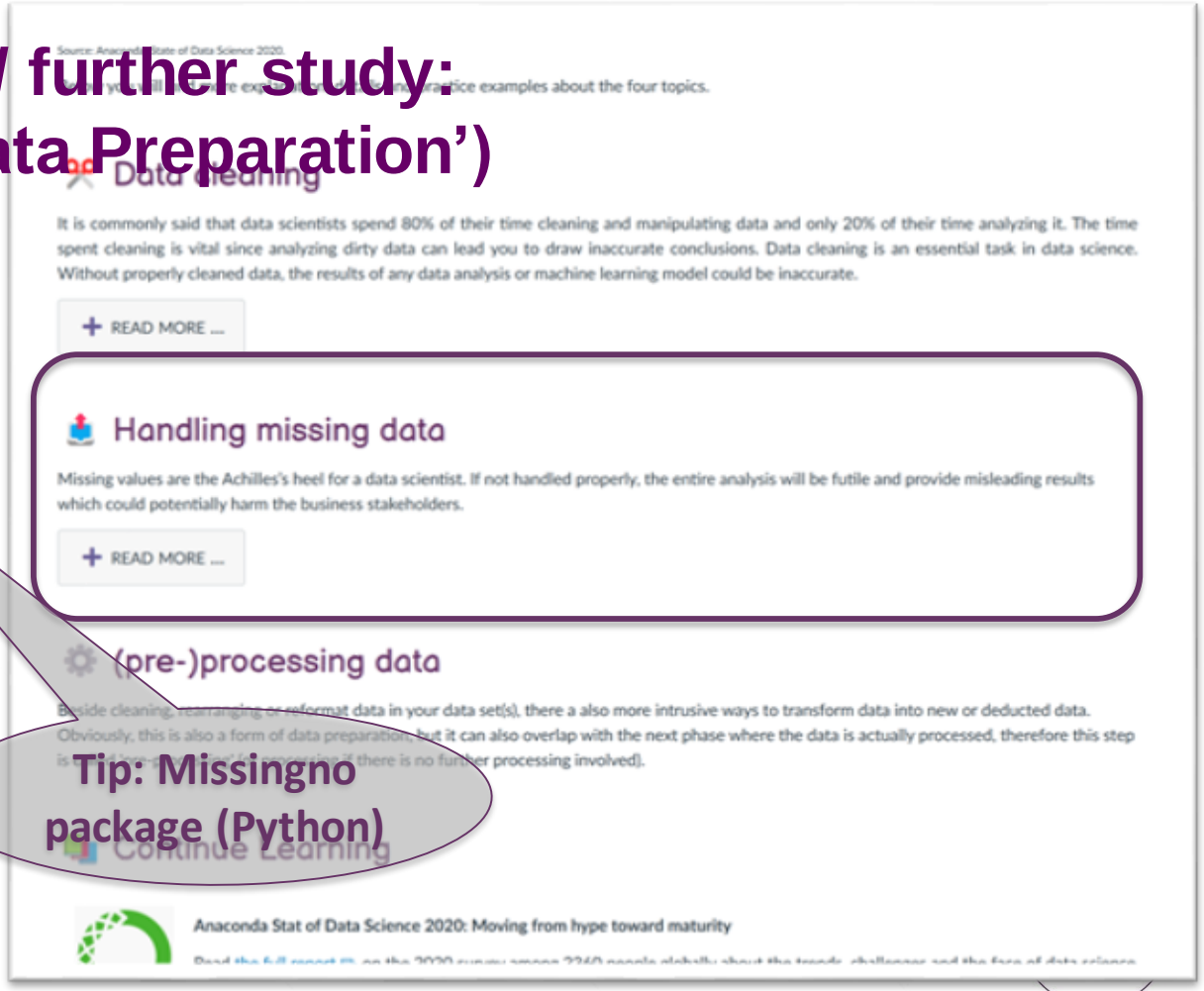
Fontys University of Applied Sciences

THINK BIGGER

More information / further study: Canvas (under ‘Data Preparation’)

Additional links/explanation:

- [article/tutorial: Missing values and using MissingNo package
- [article] How to deal with Missing Values in Machine Learning
- [article] Checking and Understanding Missing Data



Tip: Missingno package (Python)