DATA PREPARATION AND FEATURE ENGINEERING

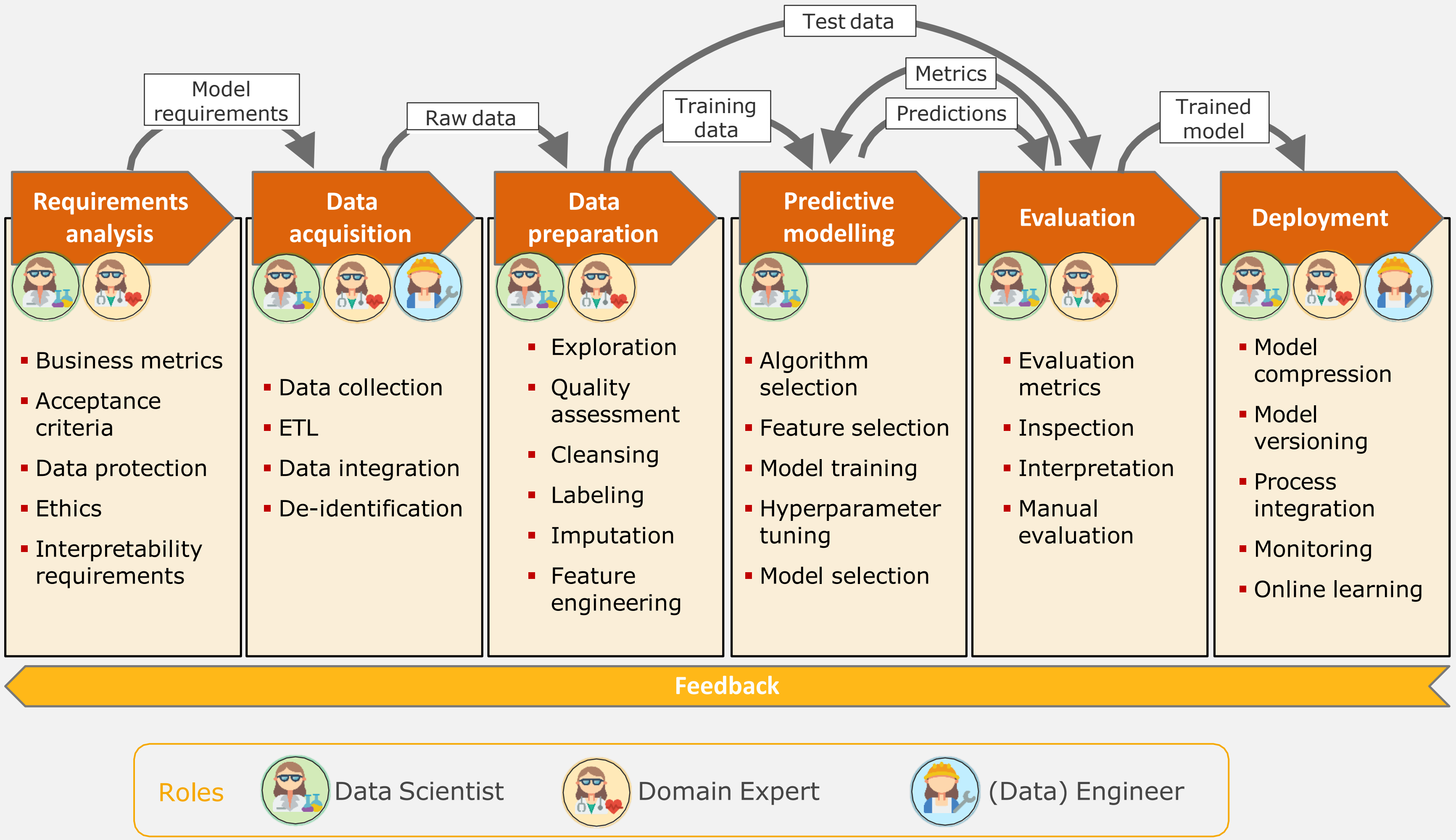
Lecture 4

MAL1, 2025

DATA PREPARATION AND FEATURE ENGINEERING

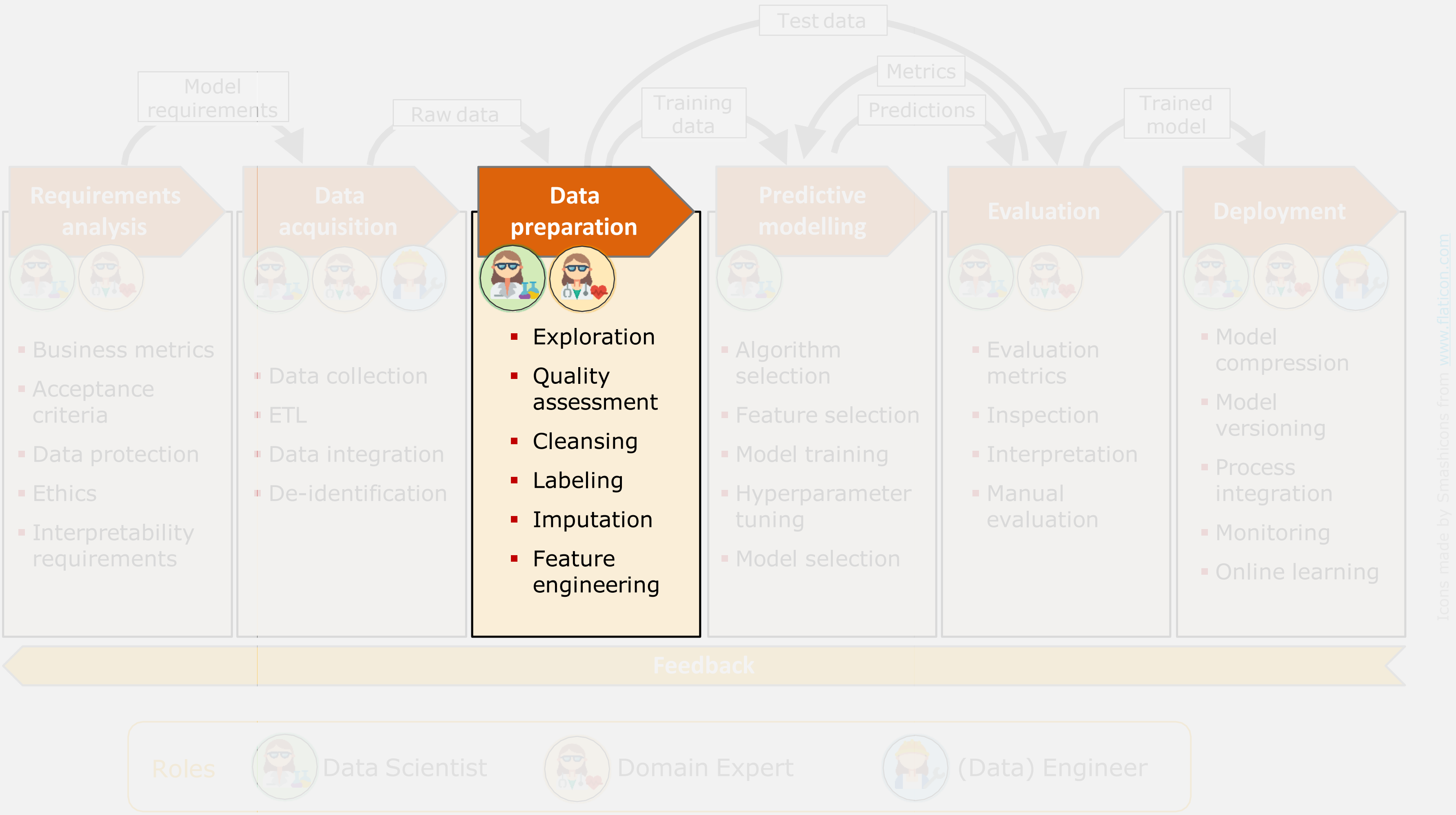
- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

A MACHINE LEARNING PROJECT

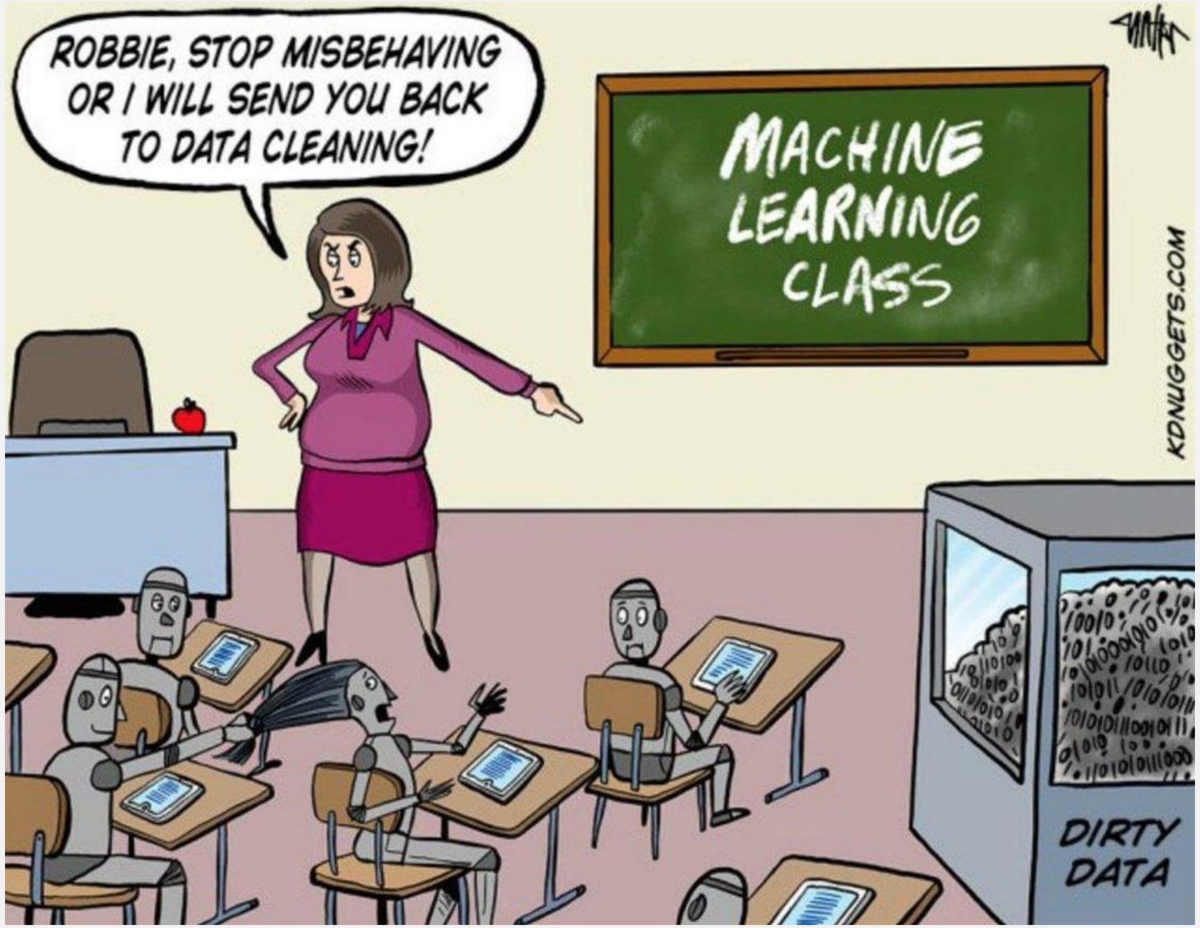


Icons made by Smashicons from www.flaticon.com

A MACHINE LEARNING PROJECT



THE IMPORTANCE OF DATA PREPARATION



Garbage in ← data

⟹

Garbage out ← model

THE TRAVELING SALESPERSONS

Salesperson ID	Years In Business	Total Sales ($)	Region	Gender	Avg Discount (%)	Customer Satisfaction	Training Hours
1	2	200000	North	Male	NaN	3.5	400
2	5	550000	NaN	Female	NaN	4.0	50
3	10	980000	West	Male	14.3	NaN	10
4	1	80000	North	Female	NaN	5.0	100
5	15	1600000	North	Male	NaN	4.5	10
6	7	900000	East	Female	NaN	4.2	5
7	20	2100000	South	Male	10.1	2.5	200

this data set is not ready

THE TRAVELING SALESPERSONS

Years In Business	Total Sales ($)	Region	Gender	Avg Discount (%)	Customer Satisfaction	Training Hours
2	200000	North	Male	NaN	3.5	400
5	550000	NaN	Female	NaN	4.0	50
10	980000	West	Male	14.3	NaN	10
1	80000	North	Female	NaN	5.0	100
15	1600000	North	Male	NaN	4.5	10
7	900000	East	Female	NaN	4.2	5
20	2100000	South	Male	10.1	2.5	200

Feature selection: get rid of sales-person ID

DATA PREPARATION AND FEATURE ENGINEERING

- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

MISSING VALUES

Years In Business	Total Sales ($)	Region	Gender	Avg Discount (%)	Customer Satisfaction	Training Hours
2	200000	North	Male	NaN	3.5	400
5	550000	NaN	Female	NaN	4.0	50
10	980000	West	Male	14.3	NaN	10
1	80000	North	Female	NaN	5.0	100
15	1600000	North	Male	NaN	4.5	10
7	900000	East	Female	NaN	4.2	5
20	2100000	South	Male	10.1	2.5	200

if a feature is mostly NaNs,
get rid of it (~>20%)

MISSING VALUES

Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	Training Hours
2	200000	North	Male	3.5	400
5	550000	NaN	Female	4.0	50
10	980000	West	Male	NaN	10
1	80000	North	Female	5.0	100
15	1600000	North	Male	4.5	10
7	900000	East	Female	4.2	5
20	2100000	South	Male	2.5	200

Impute (replace)
with a likely value
- numeric: mean
- categorical: mode

MISSING VALUES

Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	Training Hours
2	200000	North	Male	3.5	400
5	550000	North	Female	4.0	50
10	980000	West	Male	3.95	10
1	80000	North	Female	5.0	100
15	1600000	North	Male	4.5	10
7	900000	East	Female	4.2	5
20	2100000	South	Male	2.5	200

STRATEGIES FOR MISSING VALUES

Note: There are no "good" ways

- Delete
  - Delete column
  - Delete row
- Impute
  - Categorical
    - Mode
    - "Missing" category
    - kNN or some other classifier
    - MICE
  - Numerical
    - Mean, median, mode
    - 0
    - kNN/linear regression
    - MICE

```
from sklearn.impute import KNNImputer
```

MICE: MULTIPLE IMPUTATIONS BY CHAINED EQUATIONS

A	B	C
	4.2	7.8
3.1	3.1	
4.3		6.3
9.8	5.5	8.1

impute with mean

A	B	C
5.7	4.2	7.8
3.1	3.1	7.4
4.3	4.3	6.3
9.8	5.5	8.1

A back to missing

A	B	C
	4.2	7.8
3.1	3.1	7.4
4.3	4.3	6.3
9.8	5.5	8.1

linear regression with A as target

A	B	C
6.3	4.2	7.8
3.1	3.1	7.4
4.3	4.3	6.3
9.8	5.5	8.1

B back to missing

A	B	C
6.3	4.2	7.8
3.1	3.1	7.4
4.3		6.3
9.8	5.5	8.1

linear regression with B as target

A	B	C
6.3	4.2	7.8
3.1	3.1	7.4
4.3	4.4	6.3
9.8	5.5	8.1

C back to missing

and so on

WHY IS DATA MISSING?

Missing Not At Random
MNAR

Probability of missing X depends on the value of X

People w/ large alcohol intake less likely to report alcohol intake

⇓ BAD!

we need more data

Missing At Random
MAR

Probability of missing X does not depend on the value of X, but may depends on other features

Older people less likely to report alcohol intake

⇓ OK

Derived imputation (regression)

Missing Completely At Random
MCAR

Probability of missing X does not depend on any features at all

Inadvertent skipping of question

⇓

Simple or derived imputation

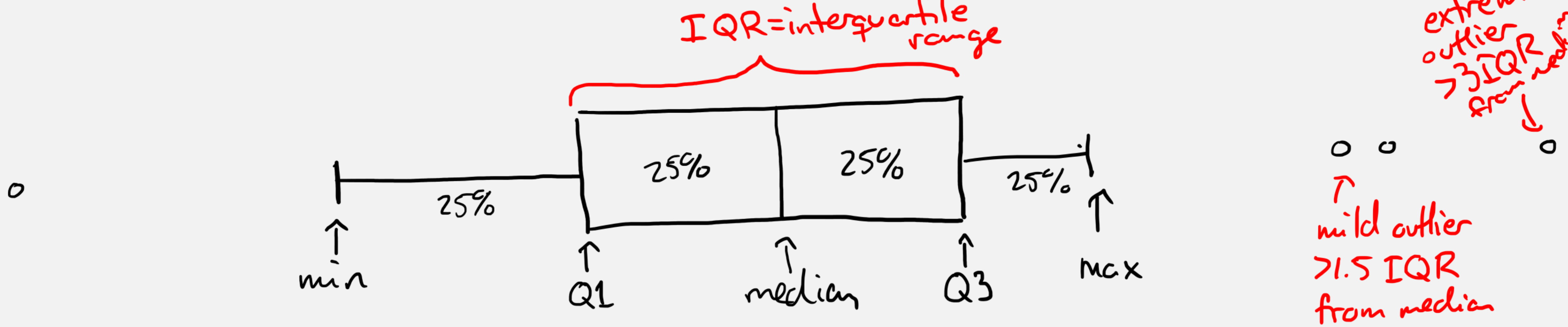
DATA PREPARATION AND FEATURE ENGINEERING

- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

OUTLIERS

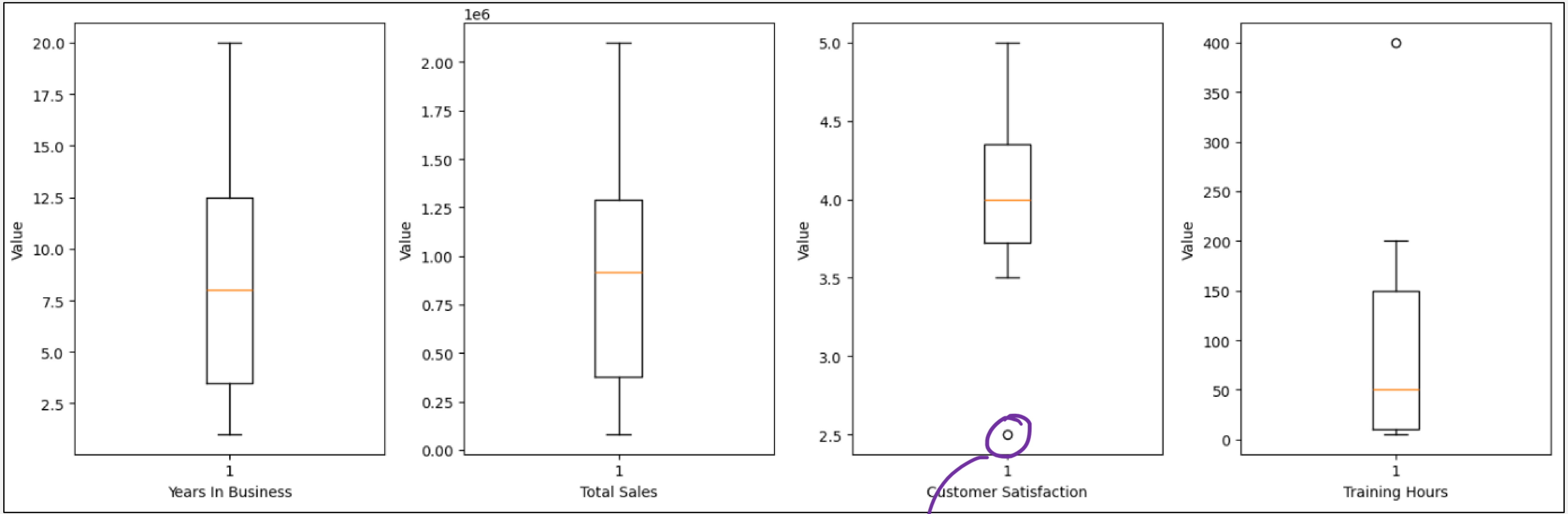
→ values outside the normal range

→ Identify w/ boxplots



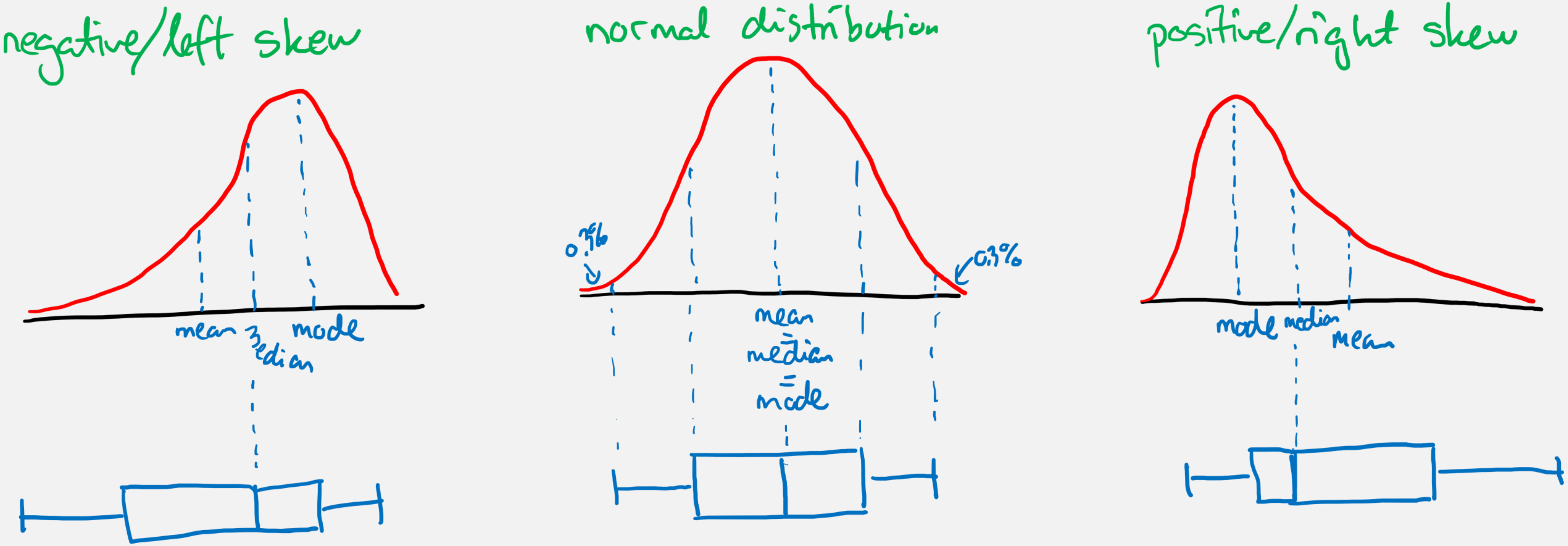
Strategies: Keep them ← if mild
Delete them } do you think it reflects something real?
Impute them }
Transform data ← if outliers come from skewedness

OUTLIERS

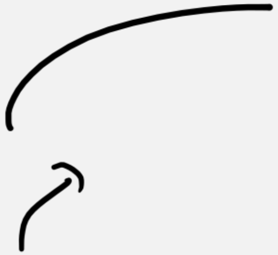


keep this

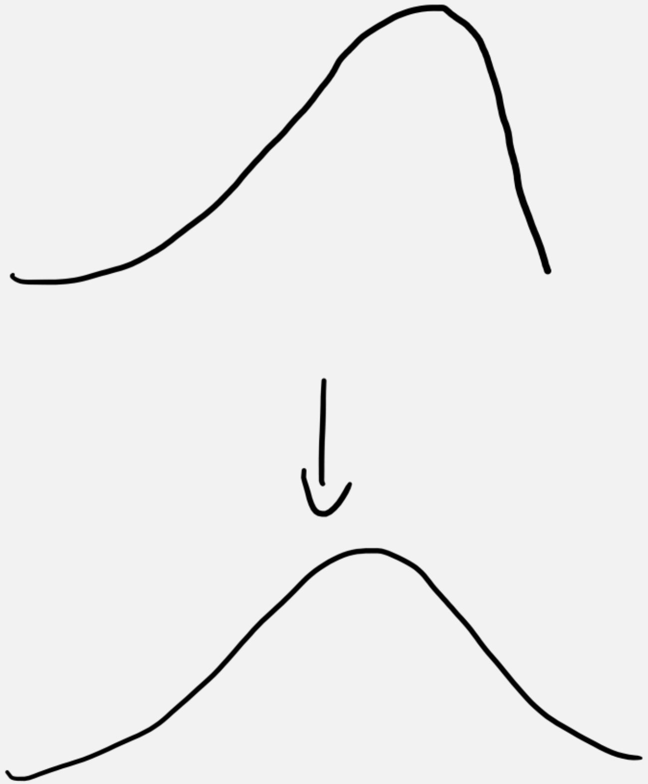
TRANSFORMING SKEWED DATA



TRANSFORMING SKEWED DATA



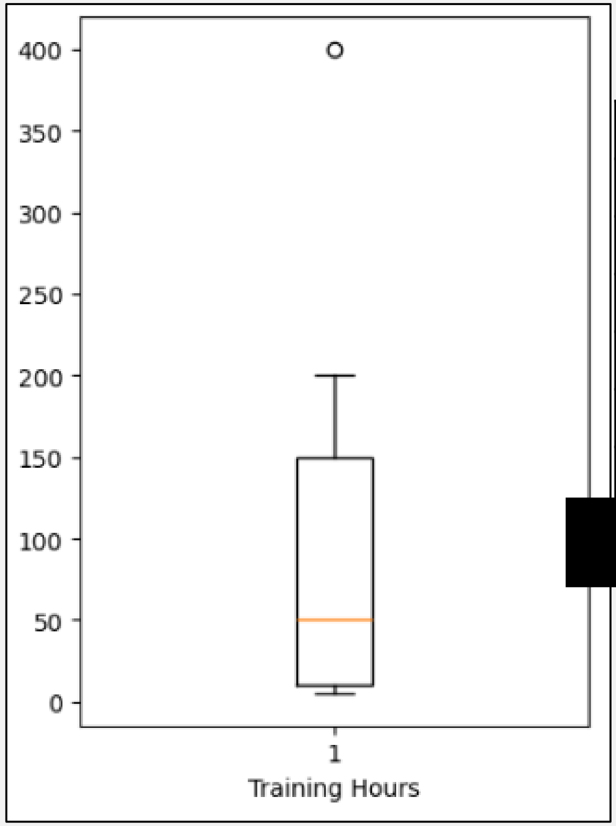
Transform left-skewed data with e^x or x^2



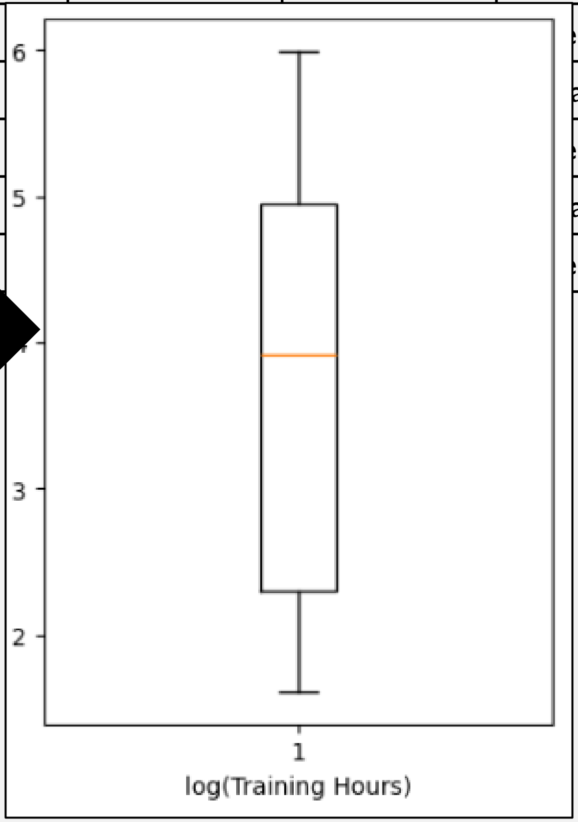
Transform right-skewed data with $\log(x)$ or $\sqrt{x}$

differences between large values are compressed by log

TRANSFORMING SKEWED DATA



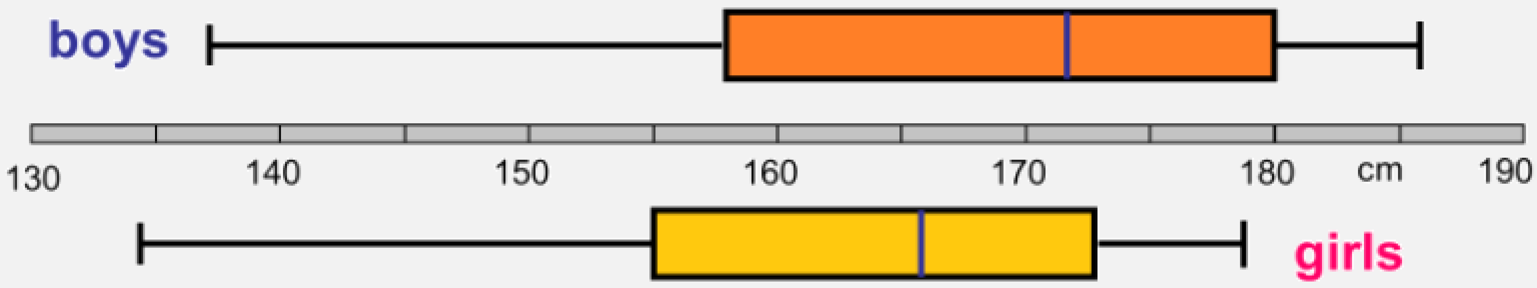
log



Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
2	200000	North	Male	3.5	5.99
5	550000	North	Female	4.0	3.91
10			e	3.95	2.30
1			ale	5.0	4.61
15			e	4.5	2.30
7			ale	4.2	1.61
20			e	2.5	5.30

Such data transformations may be relevant for skewed data even if there are no outliers!

THE BOXPLOT QUIZ



True or False?

1. Girls are, on average, taller. False
2. Girls have a higher spread. False
3. The shortest person is a girl. True
4. The tallest person is a boy. True
5. Both datasets are left skewed. True
6. The average height of boys is 172 cm. False
7. Half of the girls are between 155 and 170 cm. False
8. The average height of boys is less than the median height. True
9. Exactly half of the boys are shorter than 172 cm. True
10. Exactly half the girls are taller than 165 cm. False
11. Exactly ¾ of the girls are taller than 155 cm. True
12. Exactly ¾ of the boys are shorter than 180 cm. True
13. The population displayed is ethnic Danish. not enough info

DATA PREPARATION AND FEATURE ENGINEERING

- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

SCALING

Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
2	200000	North	Male	3.5	5.99
5	550000	North	Female	4.0	3.91
10	980000	West	Male	3.95	2.30
1	80000	North	Female	5.0	4.61
15	1600000	North	Male	4.5	2.30
7	900000	East	Female	4.2	1.61
20	2100000	South	Male	2.5	5.30

many algorithms (kNN, Ridge, lasso) depend on the scale of data

DIFFERENT TYPES OF SCALING

ideal when there are fixed boundaries

```python
from sklearn.preprocessing import MinMaxScaler
```

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

all values between 0 and 1

```python
from sklearn.preprocessing import StandardScaler
```

$$x' = \frac{x - \bar{x}}{\sigma}$$

subtract mean

divide by standard deviation

ideal when data is normally distributed

the scaled feature has $\bar{x}' = 0$ and $\sigma = 1$

SCALING

Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
-0.89	-1.06	North	Male	-0.61	1.46
-0.45	-0.54	North	Female	0.07	0.13
0.30	0.09	West	Male	0.00	-0.91
-1.04	-1.23	North	Female	1.42	0.57
1.04	1.01	North	Male	0.75	-0.91
-0.74	-0.02	East	Female	0.34	-1.35
1.79	1.74	South	Male	-1.97	1.02

DATA PREPARATION AND FEATURE ENGINEERING

- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

DEALING WITH STRINGS

Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
-1.05	-1.06	North	Male	-0.61	1.46
-0.58	-0.54	North	Female	0.07	0.13
0.20	0.09	West	Male	0.00	-0.91
-1.20	-1.23	North	Female	1.42	0.57
0.98	1.01	North	Male	0.75	-0.91
-0.11	-0.02	East	Female	0.34	-1.35
1.76	1.74	South	Male	-1.97	1.02

WHAT MAY STRINGS REPRESENT?

Text

↳ Bag of Words / CountVectorizer

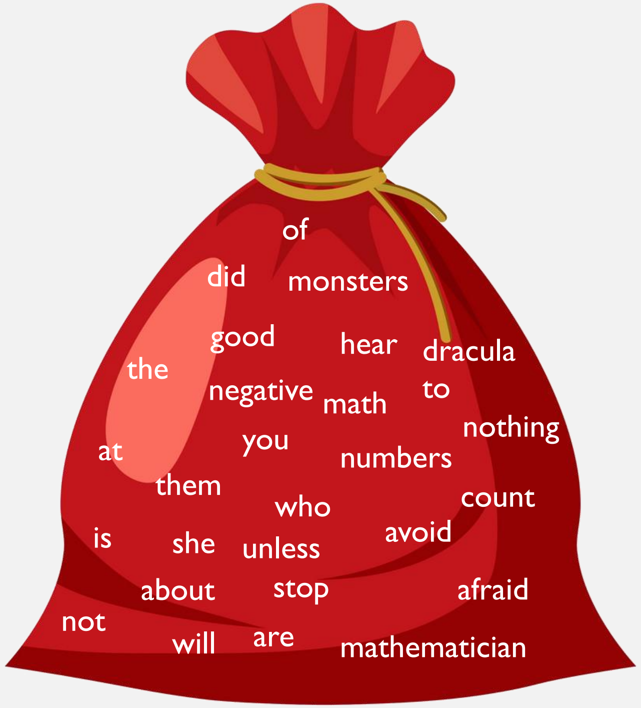
Category

↳ One-hot encoding / Dummy encoding

BAG OF WORDS

Did you hear about the mathematician who is afraid of the negative numbers? She will stop at nothing to avoid them.

Are monsters good at math? Not unless you Count Dracula.



about		1	0
afraid		1	0
are		0	1
at		1	1
avoid		1	0
count		0	1
did		1	0
dracula		0	1
good		0	1
hear		1	0
is		1	0
math		0	1
mathematician		1	0
monsters		0	1
negative		1	0
not		0	1
nothing		1	0
numbers		1	0
of		1	0
she		1	0
stop		1	0
the		2	0
them		1	0
to		1	0
unless		0	1
who		1	0
will		1	0
you		0	1

```
from sklearn.feature_extraction.text import CountVectorizer
```

ONE-HOT ENCODING

Region
North
North
West
North
North
East
South

Region North	Region West	Region East	Region South
1	0	0	0
1	0	0	0
0	1	0	0
1	0	0	0
1	0	0	0
0	0	1	0
0	0	0	1

ONE-HOT ENCODING

Years In Business	Total Sales ($)	Region North	Region West	Region East	Region South	Gender Male	Customer Satisfaction	log(Training Hours)
-1.05	-1.06	1	0	0	0	1	-0.61	1.46
-0.58	-0.54	1	0	0	0	0	0.07	0.13
0.20	0.09	0	1	0	0	1	0.00	-0.91
-1.20	-1.23	1	0	0	0	0	1.42	0.57
0.98	1.01	1	0	0	0	1	0.75	-0.91
-0.11	-0.02	0	0	1	0	0	0.34	-1.35
1.76	1.74	0	0	0	1	1	-1.97	1.02

if only two categories,
a single feature suffices

DATA PREPARATION AND FEATURE ENGINEERING

- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

FEATURE ENGINEERING

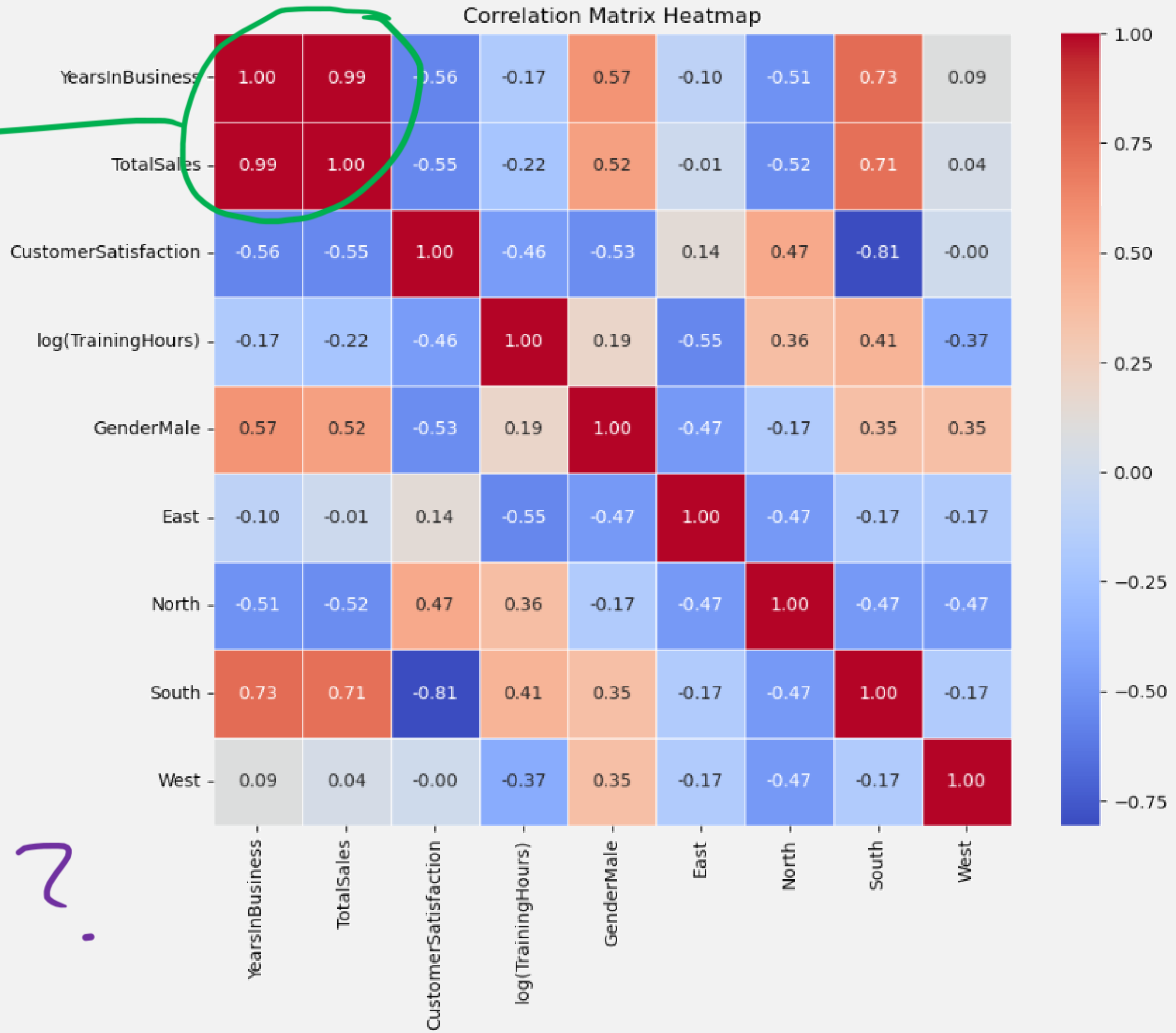
Years In Business	Total Sales ($)	Region North	Region West	Region East	Region South	Gender Male	Customer Satisfaction	log(Training Hours)
-1.05	-1.06	1	0	0	0	1	-0.61	1.46
-0.58	-0.54	1	0	0	0	0	0.07	0.13
0.20	0.09	0	1	0	0	1	0.00	-0.91
-1.20	-1.23	1	0	0	0	0	1.42	0.57
0.98	1.01	1	0	0	0	1	0.75	-0.91
-0.11	-0.02	0	0	1	0	0	0.34	-1.35
1.76	1.74	0	0	0	1	1	-1.97	1.02

Feature selection $\Leftarrow$ selecting relevant features
Feature extraction $\Leftarrow$ combining existing features

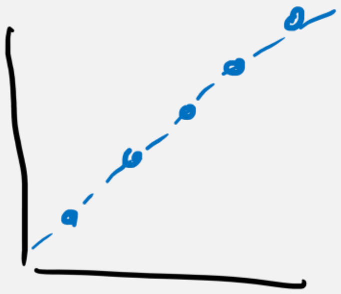
CORRELATION MATRIX

`data.corr()`

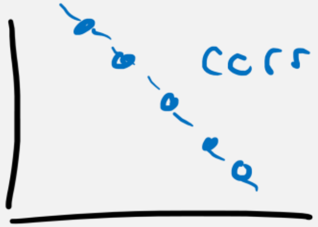
we don't want features that are highly correlated



corr = 1



corr = -1



what to do?

- delete one
- or extract new feature

Sales/Years?

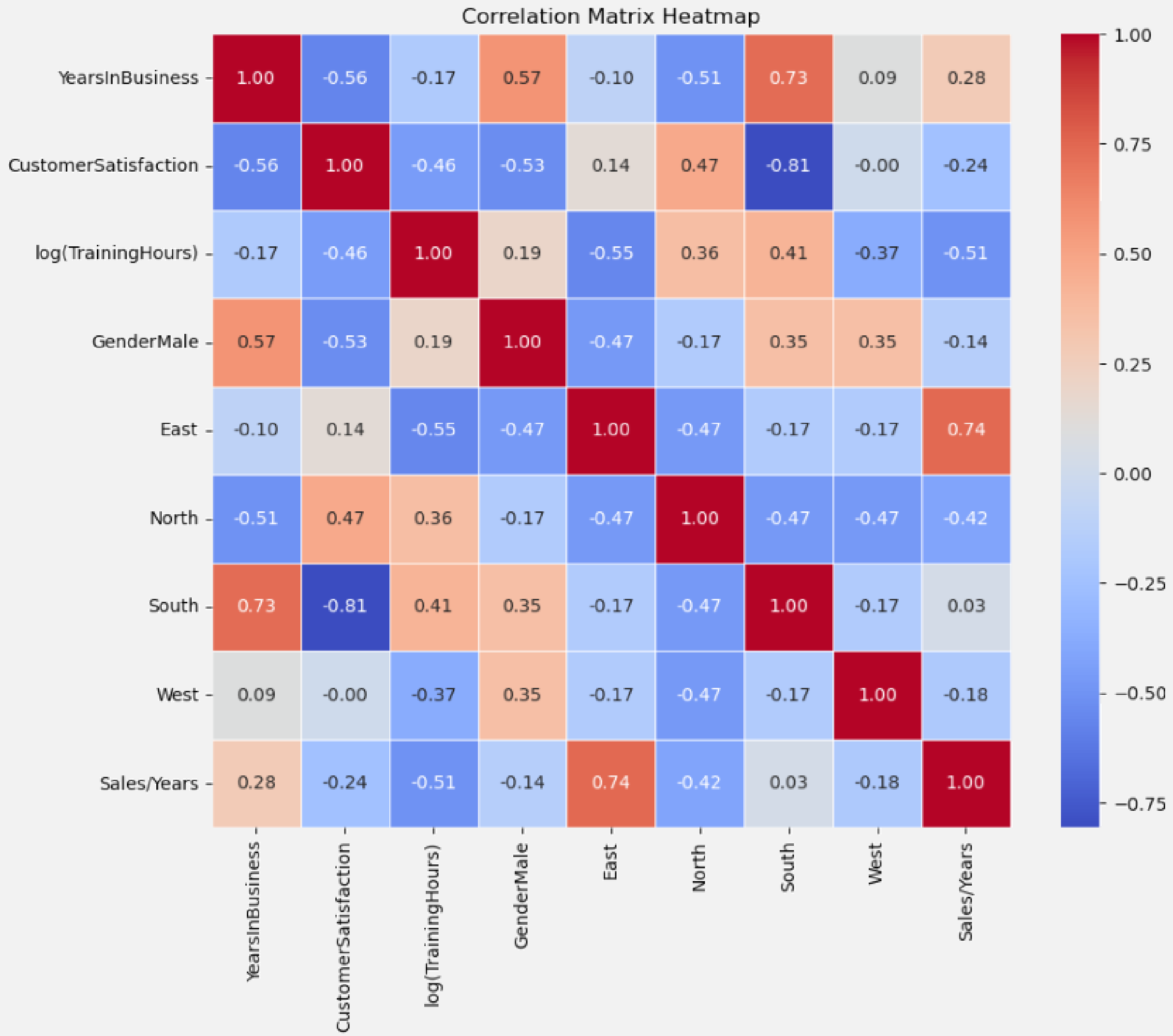
AND WITH OUR NEW FEATURE

Years In Business	Sales/Years	Region North	Region West	Region East	Region South	Gender Male	Customer Satisfaction	log(Training Hours)
-1.05	-0.30	1	0	0	0	1	-0.61	1.46
-0.58	0.44	1	0	0	0	0	0.07	0.13
0.20	-0.45	0	1	0	0	1	0.00	-0.91
-1.20	-1.78	1	0	0	0	0	1.42	0.57
0.98	0.20	1	0	0	0	1	0.75	-0.91
-0.11	1.82	0	0	1	0	0	0.34	-1.35
1.76	0.07	0	0	0	1	1	-1.97	1.02

↳ calculated based on non-scaled data, then rescaled

↳ retain this, get rid of TotalSales

CORRELATION MATRIX (AGAIN)



OUR FINAL DATA MATRIX

Salesperson ID	Years In Business	Total Sales ($)	Region	Gender	Avg Discount (%)	Customer Satisfaction	Training Hours
1	2	200000	North	Male	NaN	3.5	400
2	5	550000	NaN	Female	NaN	4.0	50
3	10	980000	West	Male	14.3	NaN	10
4	1	80000	North	Female	NaN	5.0	100
5	15	1600000	North	Male	NaN	4.5	10
6	7	900000	East	Female	NaN	4.2	5
7	20	2100000	South	Male	10.1	2.5	200

Years In Business	Sales/Years	Region North	Region West	Region East	Region South	Gender Male	Customer Satisfaction	log(Training Hours)
-1.05	-0.30	1	0	0	0	1	-0.61	1.46
-0.58	0.44	1	0	0	0	0	0.07	0.13
0.20	-0.45	0	1	0	0	1	0.00	-0.91
-1.20	-1.78	1	0	0	0	0	1.42	0.57
0.98	0.20	1	0	0	0	1	0.75	-0.91
-0.11	1.82	0	0	1	0	0	0.34	-1.35
1.76	0.07	0	0	0	1	1	-1.97	1.02

- Explain why data preparation is necessary
- Explain the steps needed to prepare a dataset
- Prepare a dataset for use in ML models in sklearn