DATA PREPARATION AND FEATURE ENGINEERING

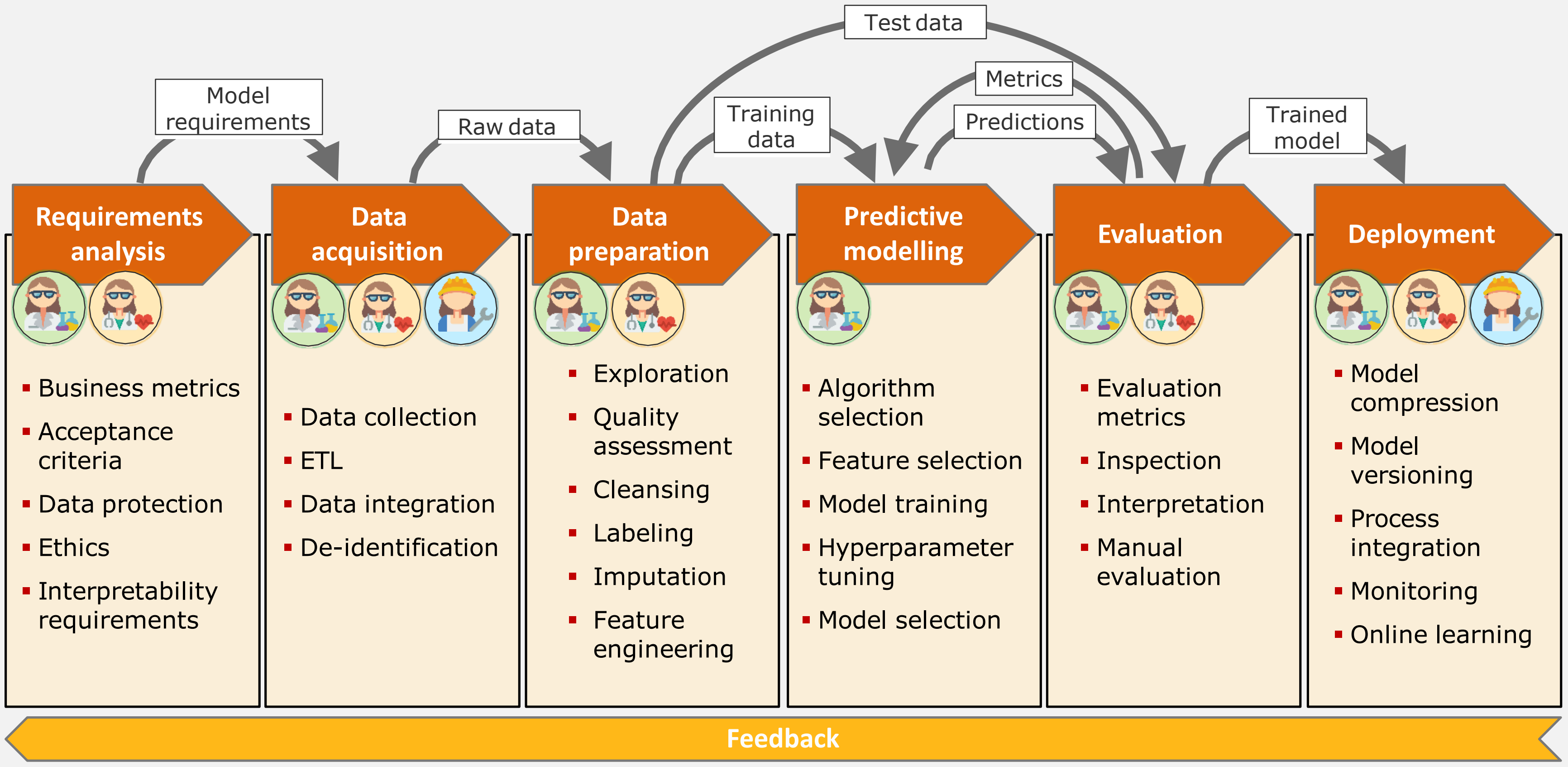
Lecture 4

MAL1, 2025

DATA PREPARATION AND FEATURE ENGINEERING

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

A MACHINE LEARNING PROJECT



Model requirements

Raw data

Training data

Test data

Metrics

Predictions

Trained model

Requirements analysis

- Business metrics
- Acceptance criteria
- Data protection
- Ethics
- Interpretability requirements

Data acquisition

- Data collection
- ETL
- Data integration
- De-identification

Data preparation

- Exploration
- Quality assessment
- Cleansing
- Labeling
- Imputation
- Feature engineering

Predictive modelling

- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Evaluation

- Evaluation metrics
- Inspection
- Interpretation
- Manual evaluation

Deployment

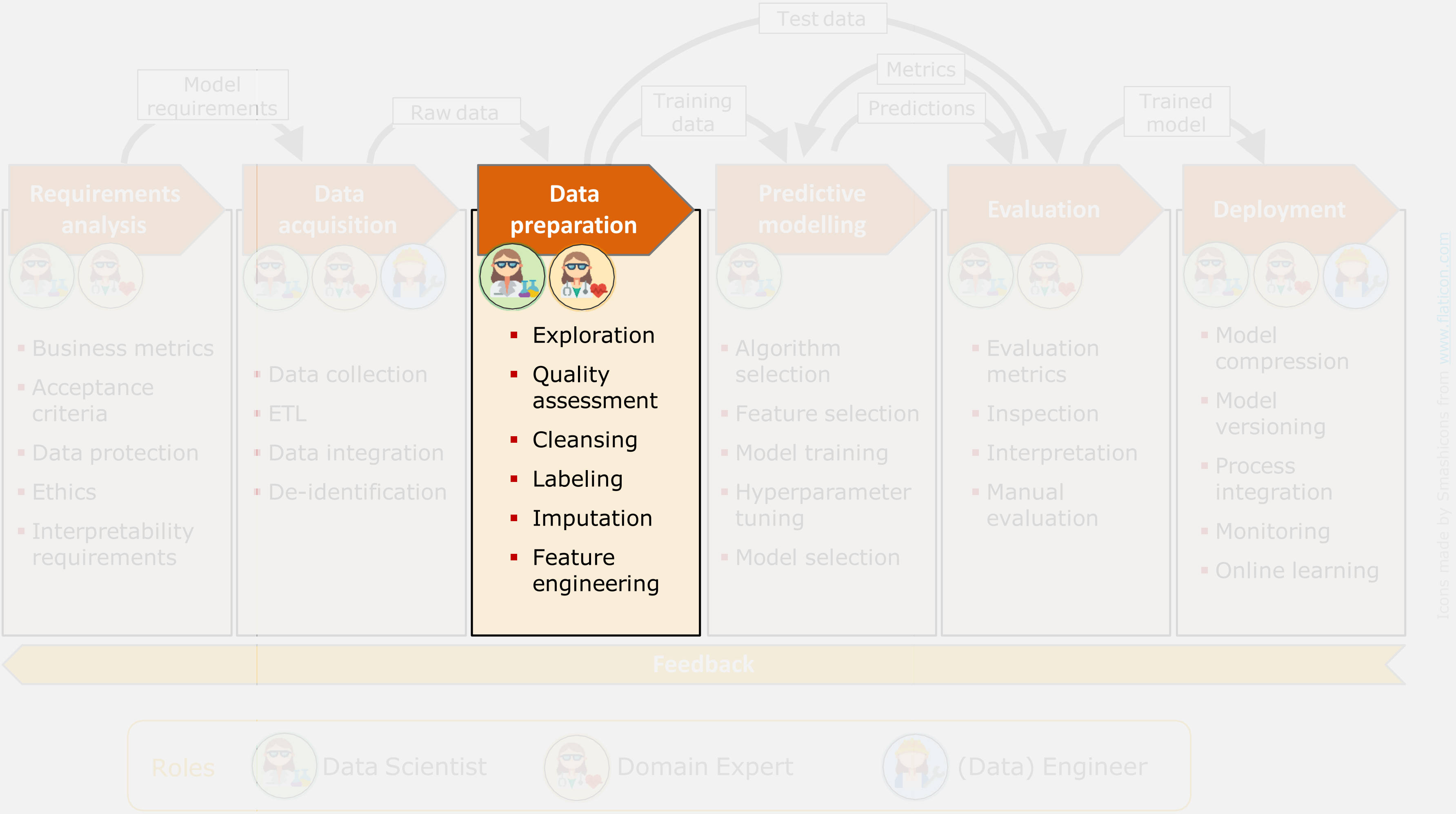
- Model compression
- Model versioning
- Process integration
- Monitoring
- Online learning

Feedback

Roles: Data Scientist, Domain Expert, (Data) Engineer

Icons made by Smashicons from www.flaticon.com

A MACHINE LEARNING PROJECT



Requirements analysis
- Business metrics
- Acceptance criteria
- Data protection
- Ethics
- Interpretability requirements

Data acquisition
- Data collection
- ETL
- Data integration
- De-identification

Data preparation
- Exploration
- Quality assessment
- Cleansing
- Labeling
- Imputation
- Feature engineering

Predictive modelling
- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

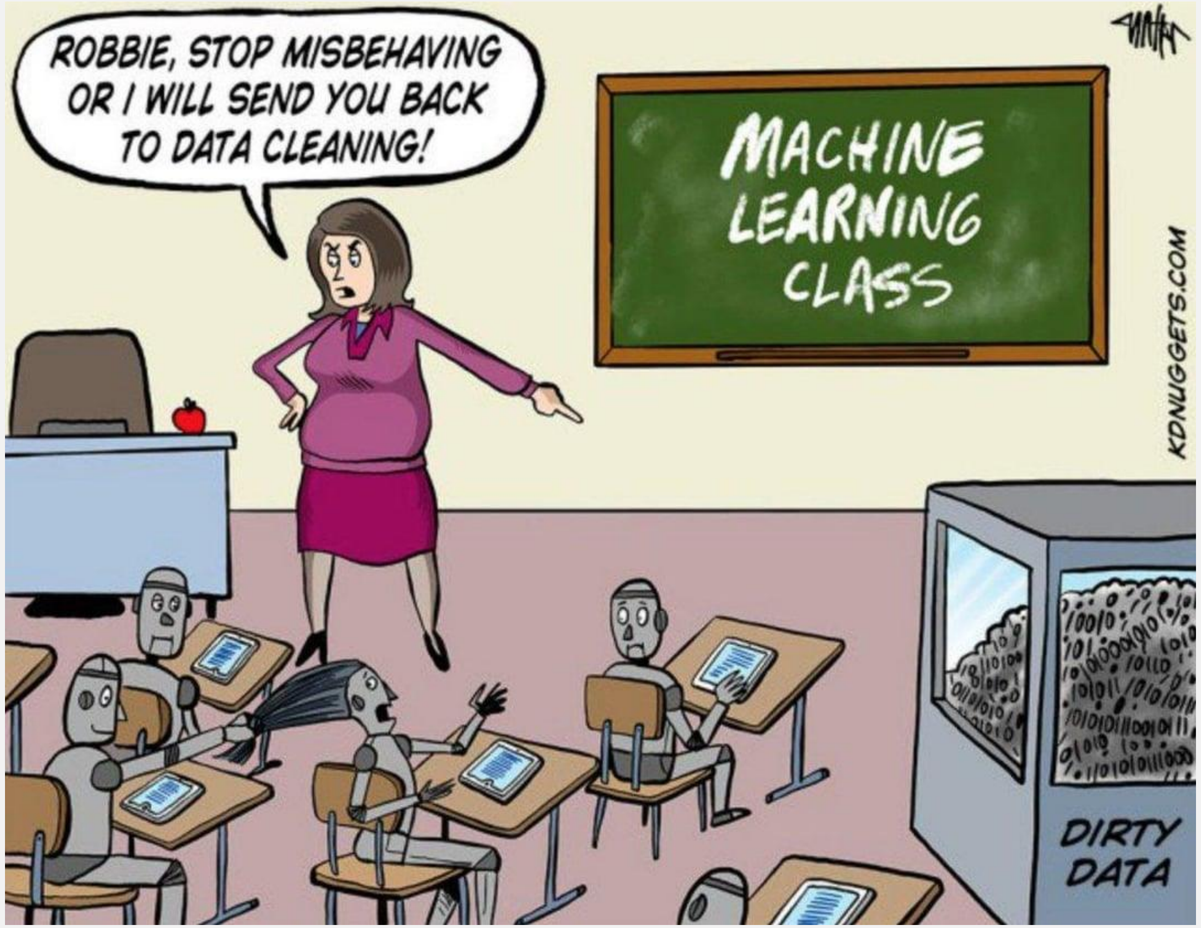
Evaluation
- Evaluation metrics
- Inspection
- Interpretation
- Manual evaluation

Deployment
- Model compression
- Model versioning
- Process integration
- Monitoring
- Online learning

Roles: Data Scientist, Domain Expert, (Data) Engineer

Icons made by Smashicons from www.flaticon.com

THE IMPORTANCE OF DATA PREPARATION


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

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

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

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

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

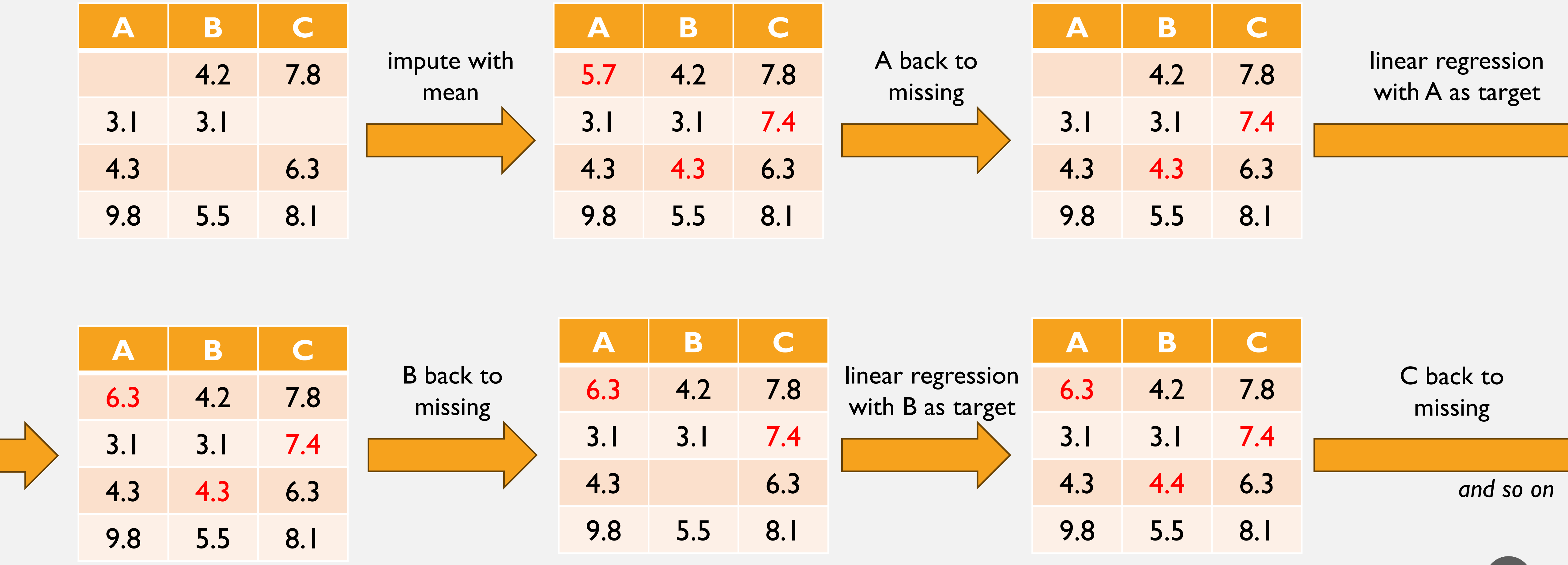
MISSING VALUES

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

```
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

Missing At Random
MAR

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

Missing Completely At Random
MCAR

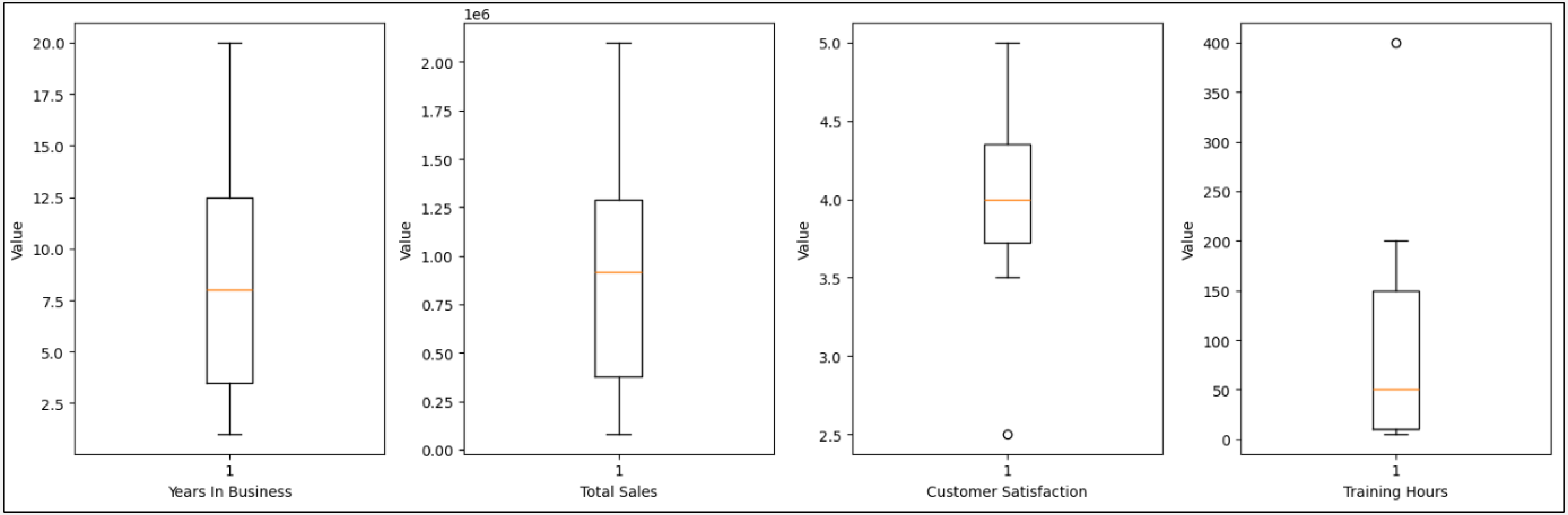
Probability of missing X does not depend on any features at all

DATA PREPARATION AND FEATURE ENGINEERING

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

OUTLIERS

OUTLIERS

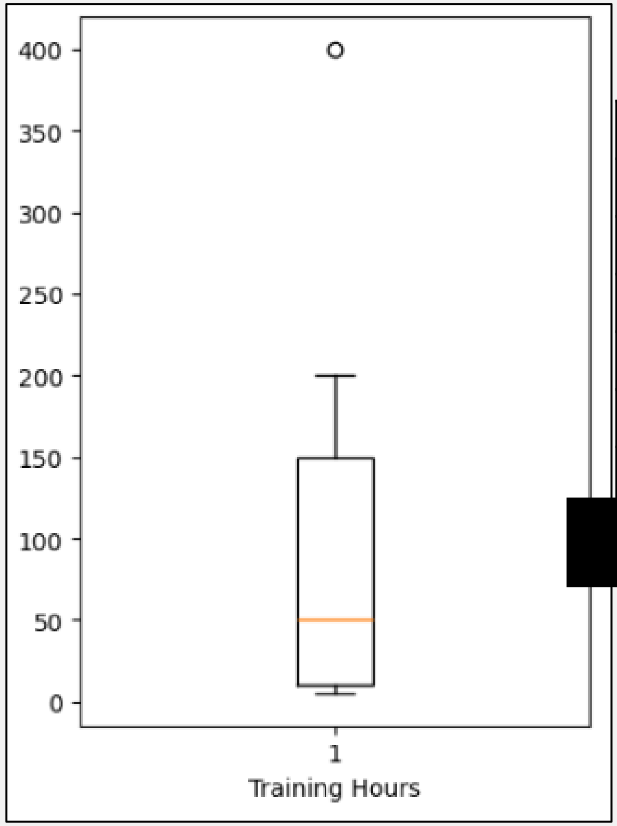

1e6
Value
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
1
Years In Business
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1
Total Sales
5.0
4.5
4.0
3.5
3.0
2.5
1
Customer Satisfaction
400
350
300
250
200
150
100
50
0
1
Training Hours

TRANSFORMING SKEWED DATA

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

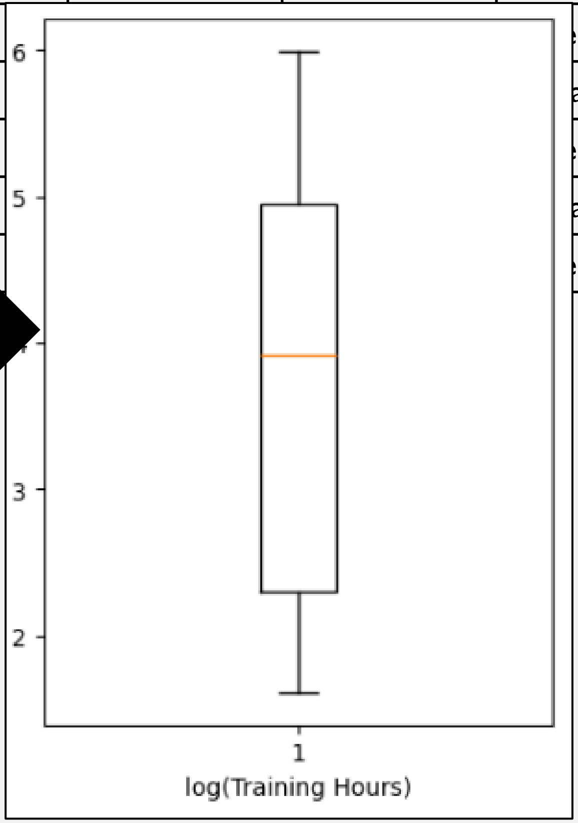
Transform right-skewed data with $\log(x)$ or $\sqrt{x}$

TRANSFORMING SKEWED DATA



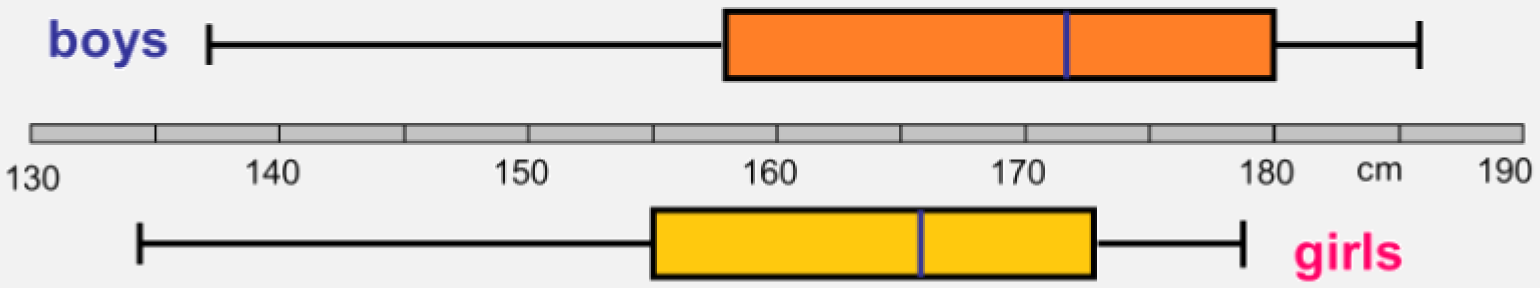
Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
2	200000	North	Male	3.5	5.99
5	550000	North	Female	4.0	3.91
10			e	3.95	2.30
1			ale	5.0	4.61
15			e	4.5	2.30
7			ale	4.2	1.61
20			e	2.5	5.30

log



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

THE BOXPLOT QUIZ



True or False?

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

DIFFERENT TYPES OF SCALING

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

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

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

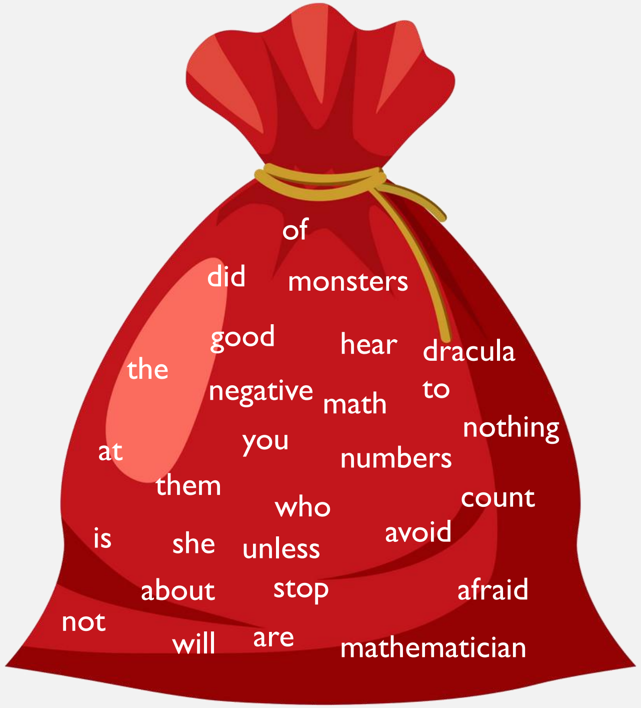
Years In Business	Total Sales ($)	Region	Gender	Customer Satisfaction	log(Training Hours)
-1.05	-1.06	North	Male	-0.61	1.46
-0.58	-0.54	North	Female	0.07	0.13
0.20	0.09	West	Male	0.00	-0.91
-1.20	-1.23	North	Female	1.42	0.57
0.98	1.01	North	Male	0.75	-0.91
-0.11	-0.02	East	Female	0.34	-1.35
1.76	1.74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

WHAT MAY STRINGS REPRESENT?

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
	0	0	0
	0	0	0
0		0	0
	0	0	0
	0	0	0
0	0		0
0	0	0	

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

DATA PREPARATION AND FEATURE ENGINEERING

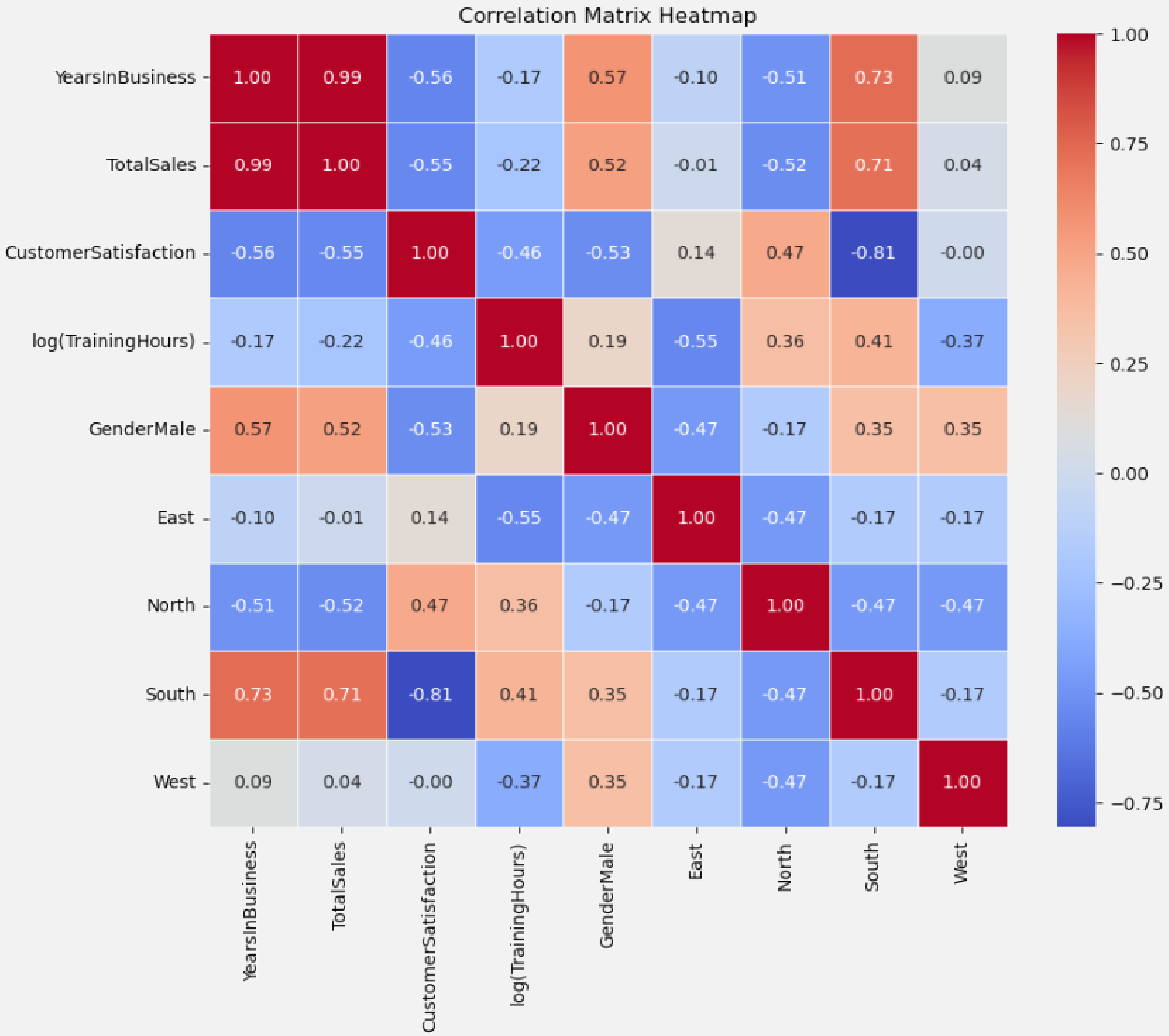
- Overview
- Missing data
- Outliers
- Scaling
- String data
- Feature engineering

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

CORRELATION MATRIX

```
data.corr()
```

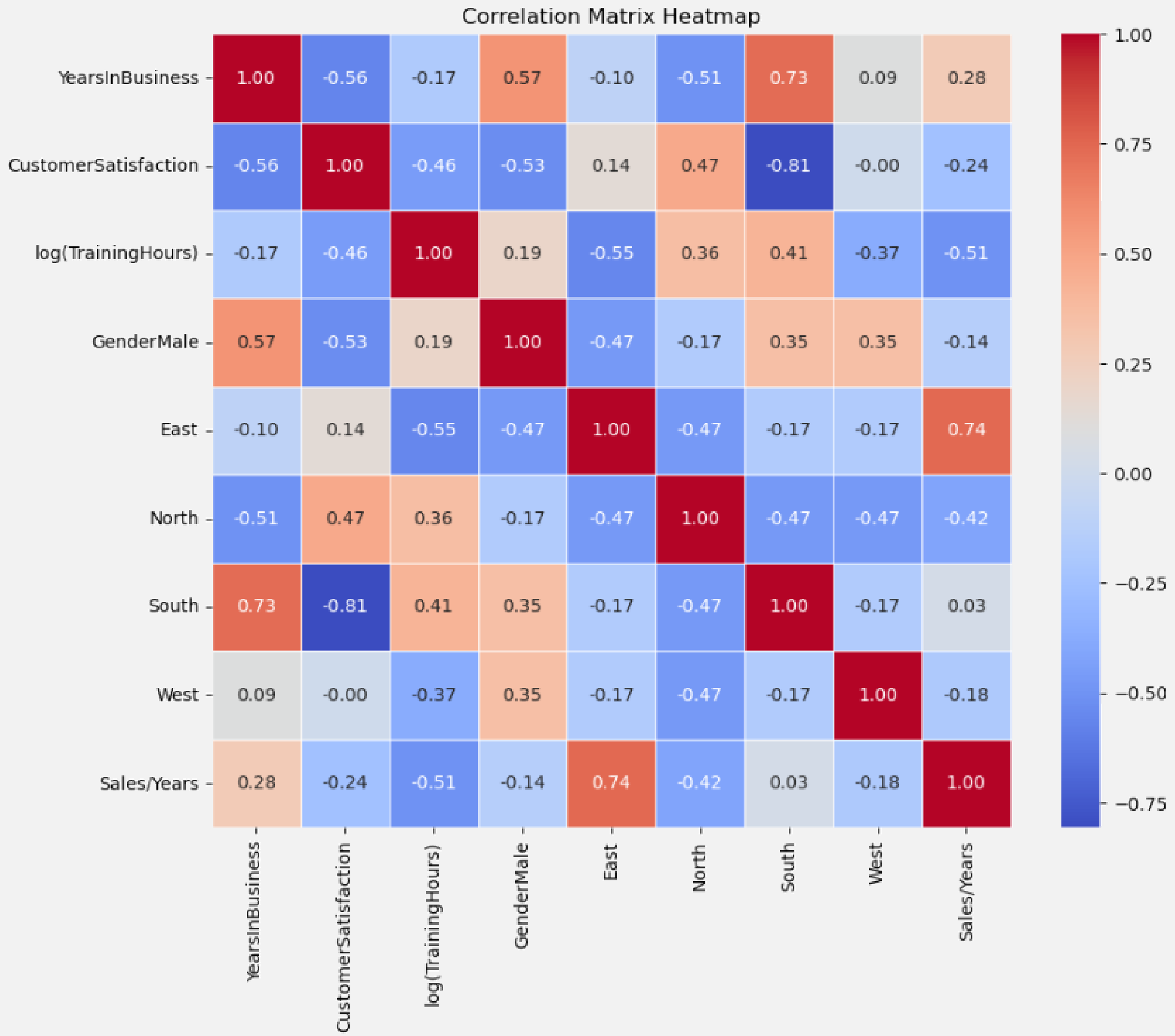


	YearsInBusiness	TotalSales	CustomerSatisfaction	log(TrainingHours)	GenderMale	East	North	South	West
YearsInBusiness	1.00	0.99	-0.56	-0.17	0.57	-0.10	-0.51	0.73	0.09
TotalSales	0.99	1.00	-0.55	-0.22	0.52	-0.01	-0.52	0.71	0.04
CustomerSatisfaction	-0.56	-0.55	1.00	-0.46	-0.53	0.14	0.47	-0.81	-0.00
log(TrainingHours)	-0.17	-0.22	-0.46	1.00	0.19	-0.55	0.36	0.41	-0.37
GenderMale	0.57	0.52	-0.53	0.19	1.00	-0.47	-0.17	0.35	0.35
East	-0.10	-0.01	0.14	-0.55	-0.47	1.00	-0.47	-0.17	-0.17
North	-0.51	-0.52	0.47	0.36	-0.17	-0.47	1.00	-0.47	-0.47
South	0.73	0.71	-0.81	0.41	0.35	-0.17	-0.47	1.00	-0.17
West	0.09	0.04	-0.00	-0.37	0.35	-0.17	-0.47	-0.17	1.00

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

CORRELATION MATRIX (AGAIN)



	YearsInBusiness	CustomerSatisfaction	log(TrainingHours)	GenderMale	East	North	South	West	Sales/Years
YearsInBusiness	1.00	-0.56	-0.17	0.57	-0.10	-0.51	0.73	0.09	0.28
CustomerSatisfaction	-0.56	1.00	-0.46	-0.53	0.14	0.47	-0.81	-0.00	-0.24
log(TrainingHours)	-0.17	-0.46	1.00	0.19	-0.55	0.36	0.41	-0.37	-0.51
GenderMale	0.57	-0.53	0.19	1.00	-0.47	-0.17	0.35	0.35	-0.14
East	-0.10	0.14	-0.55	-0.47	1.00	-0.47	-0.17	-0.17	0.74
North	-0.51	0.47	0.36	-0.17	-0.47	1.00	-0.47	-0.47	-0.42
South	0.73	-0.81	0.41	0.35	-0.17	-0.47	1.00	-0.17	0.03
West	0.09	-0.00	-0.37	0.35	-0.17	-0.47	-0.17	1.00	-0.18
Sales/Years	0.28	-0.24	-0.51	-0.14	0.74	-0.42	0.03	-0.18	1.00

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