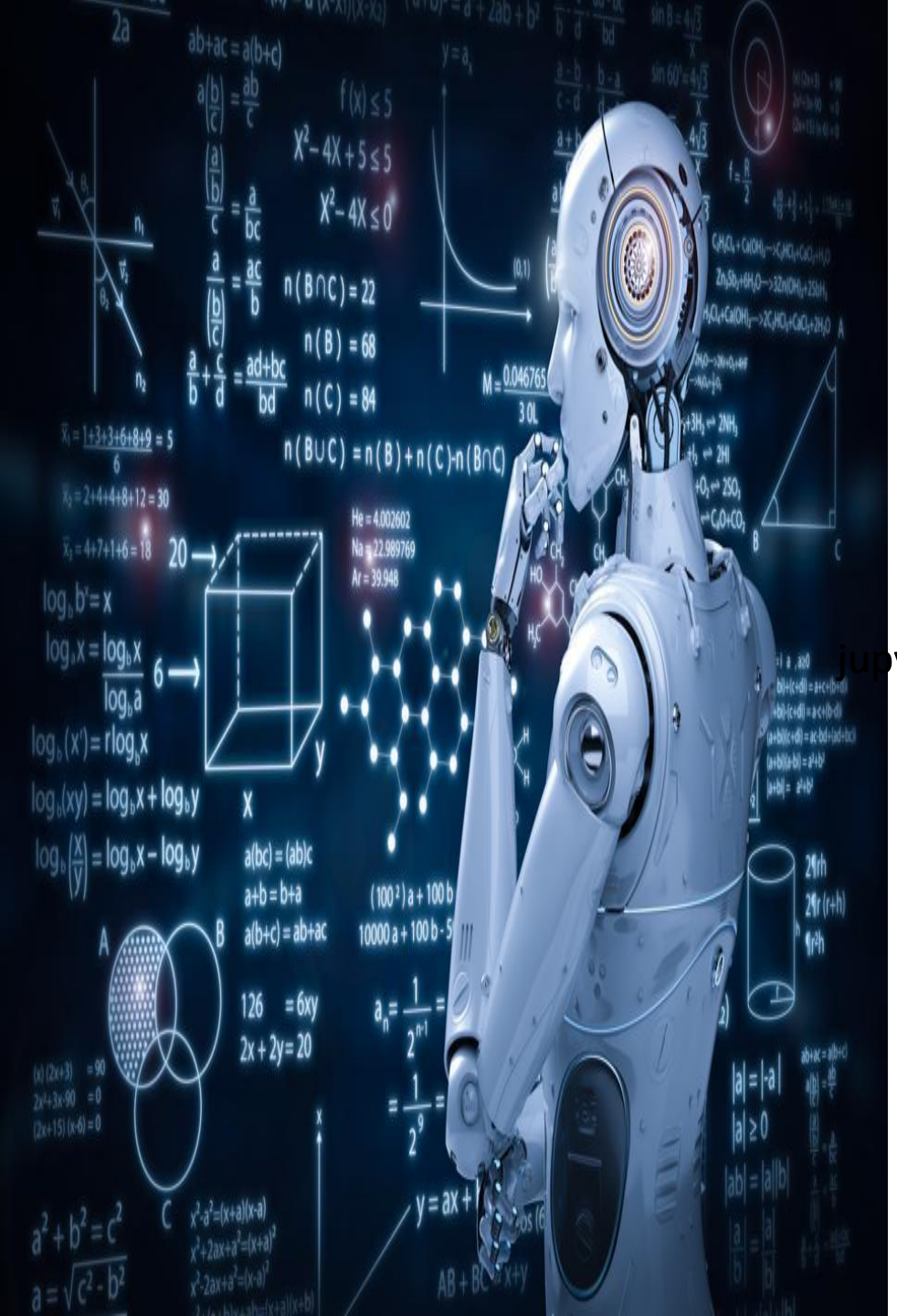


Practical Machine Learning

jupyter notebook --notebook-dir=D:/

Day 4: Sep22 DBDA

Kiran Waghmare

jupyter Day3

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code Commit

```
#Replacing the missing value
X[:,1:3]=imputer.transform(X[:,1:3])
```

In [9]:
```
X
```

Out[9]:
```
array([['France', 44.0, 72000.0],
       ['Spain', 27.0, 48000.0],
       ['Germany', 30.0, 54000.0],
       ['Spain', 38.0, 61000.0],
       ['Germany', 40.0, 63777.77777777778],
       ['France', 35.0, 58000.0],
       ['Spain', 38.77777777777778, 52000.0],
       ['France', 48.0, 79000.0],
       ['Germany', 50.0, 83000.0],
       ['France', 37.0, 67000.0]], dtype=object)
```

France :0
Germany:1
Spain :2

In [10]:
```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
label=LabelEncoder()
X[:,0]= label.fit_transform(X[:,0])
```

In [11]:
```
X
```

Out[11]:
```
array([[0, 44.0, 72000.0],
       [2, 27.0, 48000.0],
```

Imputation

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

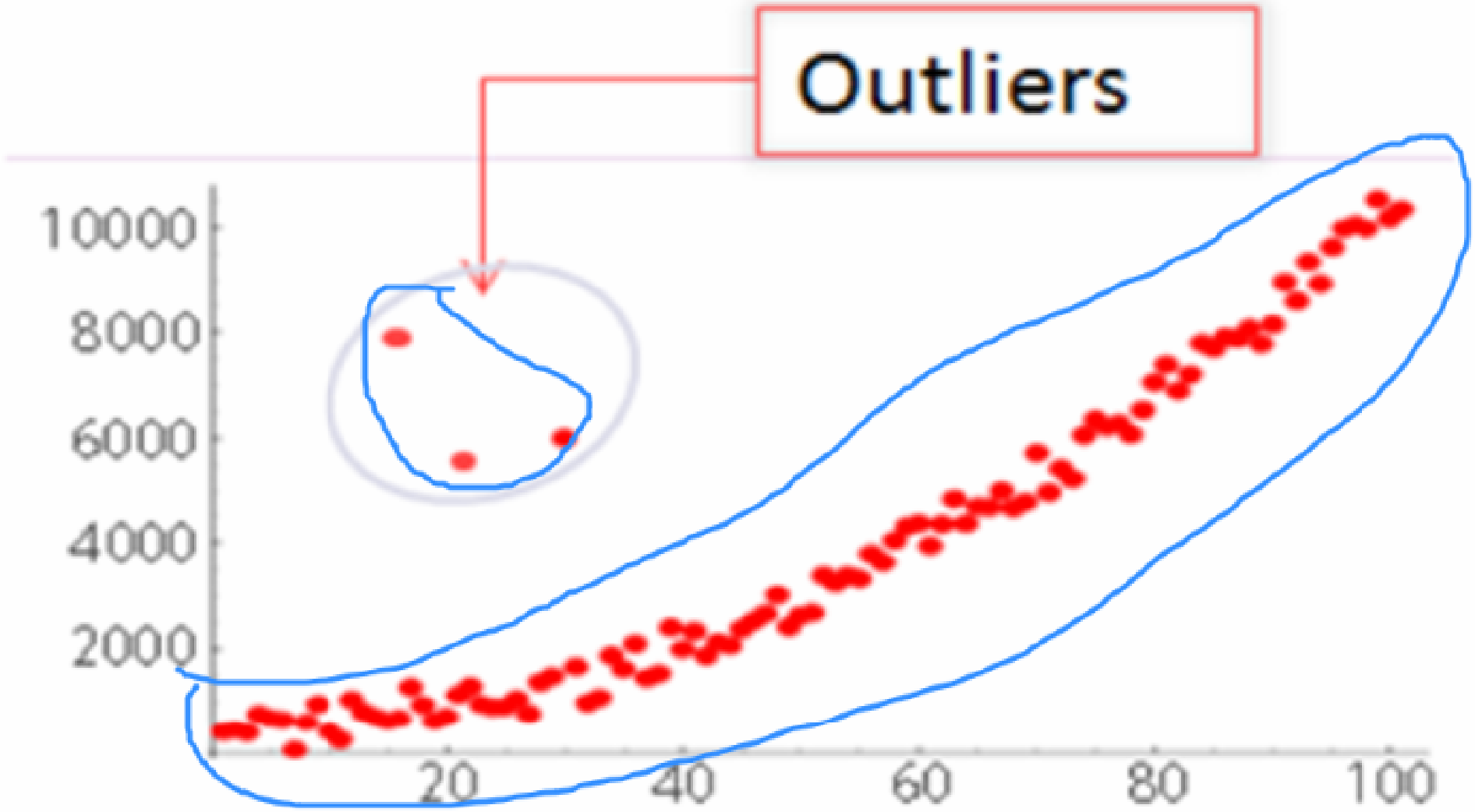
	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

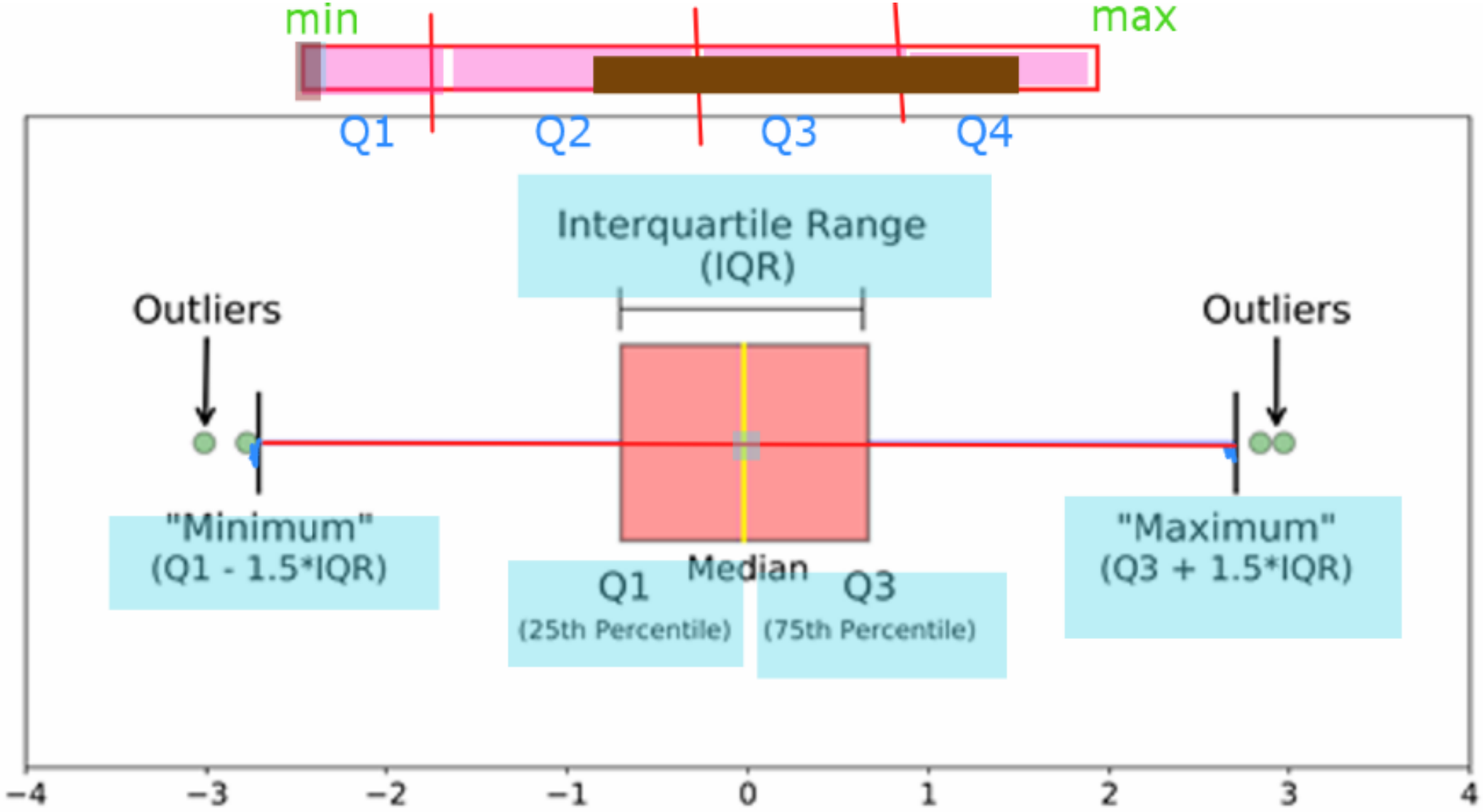
	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN

Mean Imputed

	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Handling Outliers




min
max
Q1
Q2
Q3
Q4
Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Median
Q1
(25th Percentile)
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

8.2 Transformations of Logarithmic Functions

Exponent

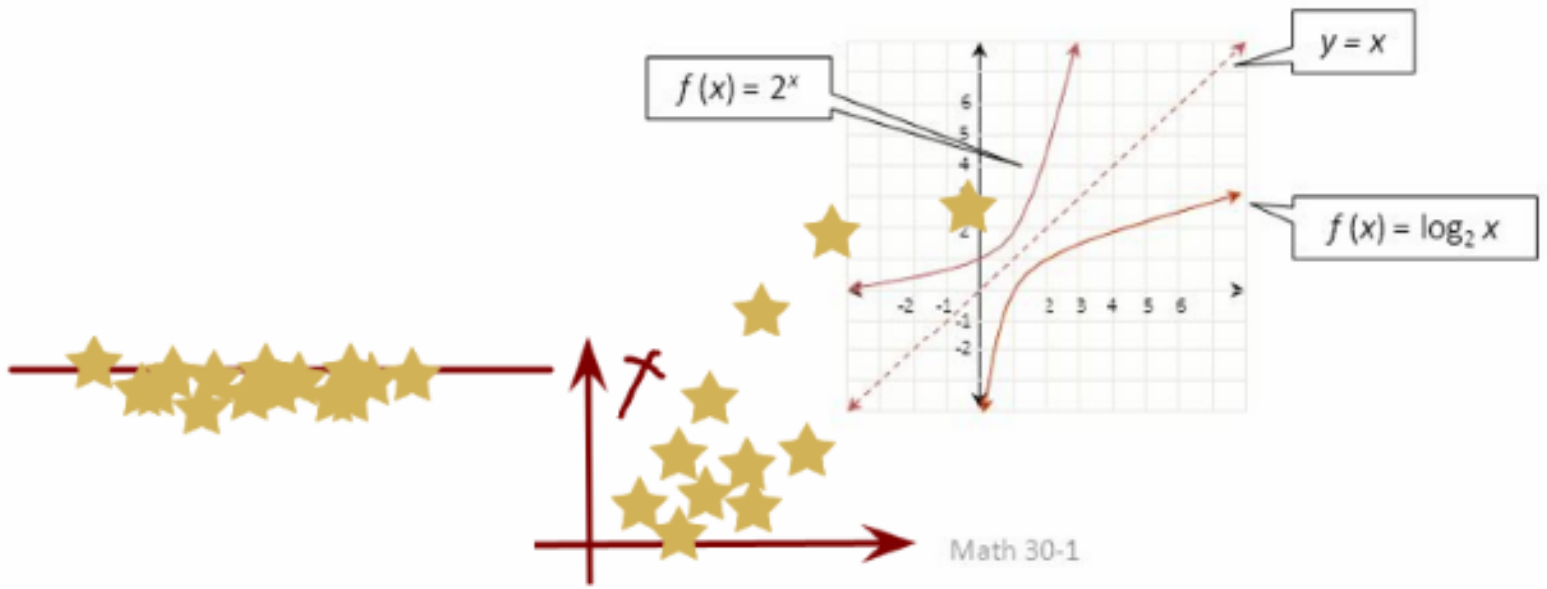
Logarithmic form: $y = \log_b x$

Base

Exponent

Exponential Form: $b^y = x$.

Base



Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Bin size=3

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

Method 1

* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

Replace with mean value

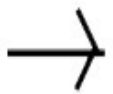
Method 2

* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Replace with nearest boundry value

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Iris dataset

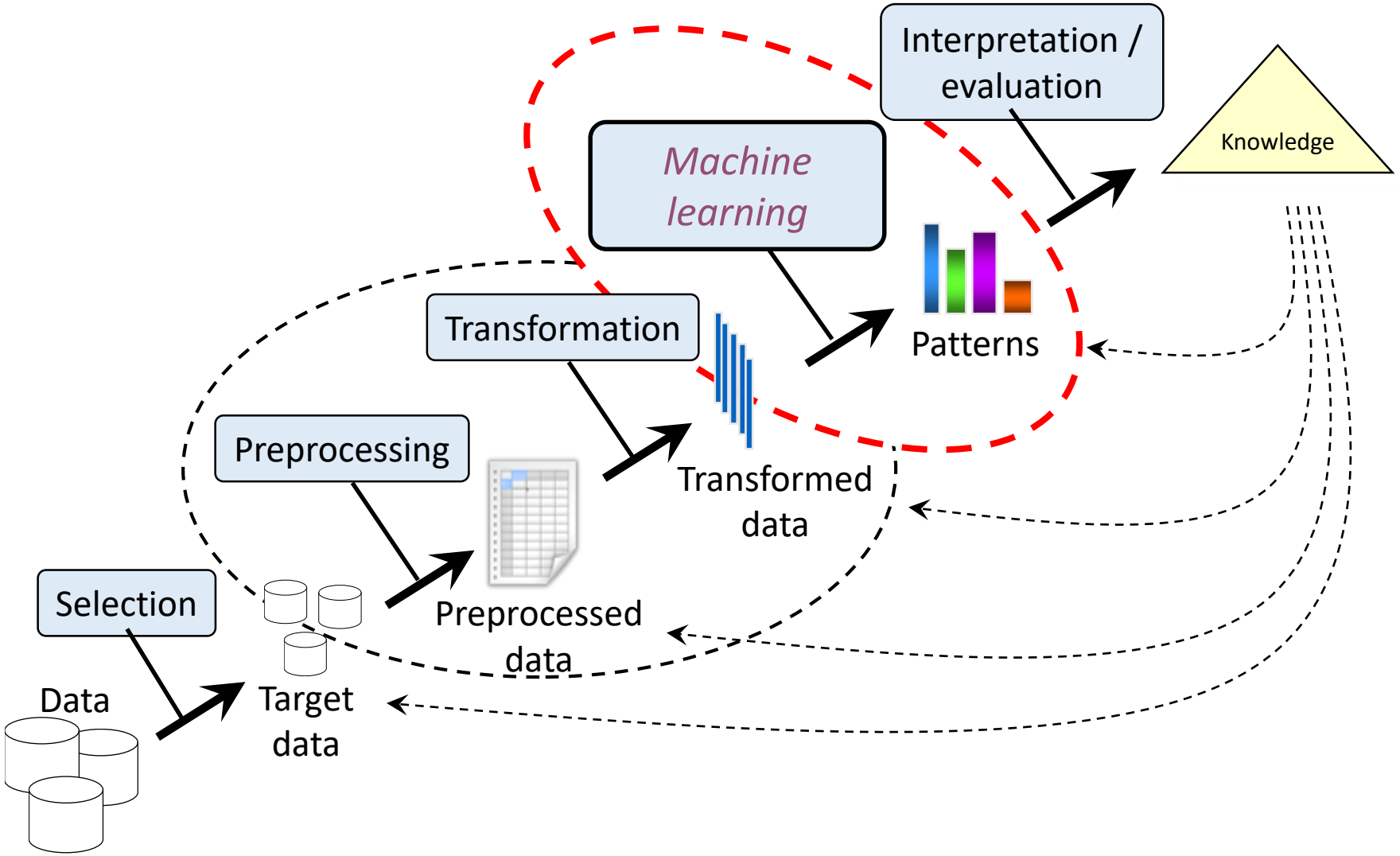
- Many exploratory data techniques are nicely illustrated with the iris dataset.
  - Dataset created by famous statistician Ronald Fisher
  - 150 samples of three species in genus *Iris* (50 each)
    - *Iris setosa*
    - *Iris versicolor*
    - *Iris virginica*
  - Four attributes
    - sepal width
    - sepal length
    - petal width
    - petal length
  - Species is class label

Iris virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

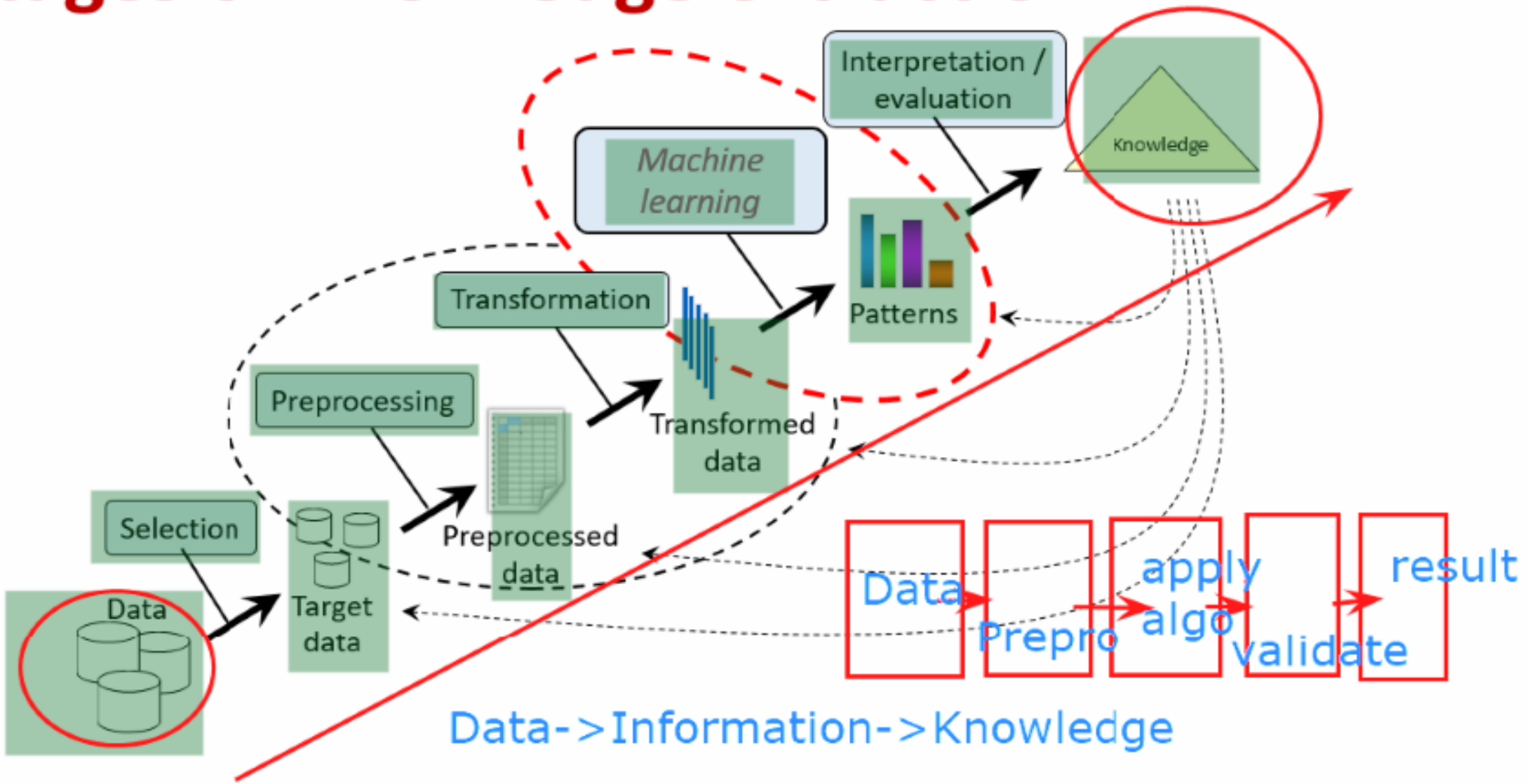
Agenda

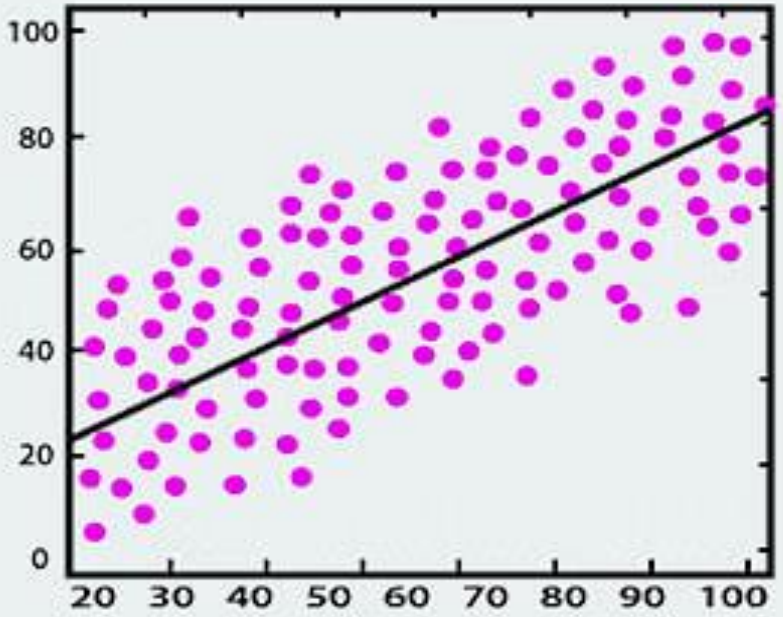
- Knowledge Extraction
- Regression Analysis

Stages of knowledge extraction



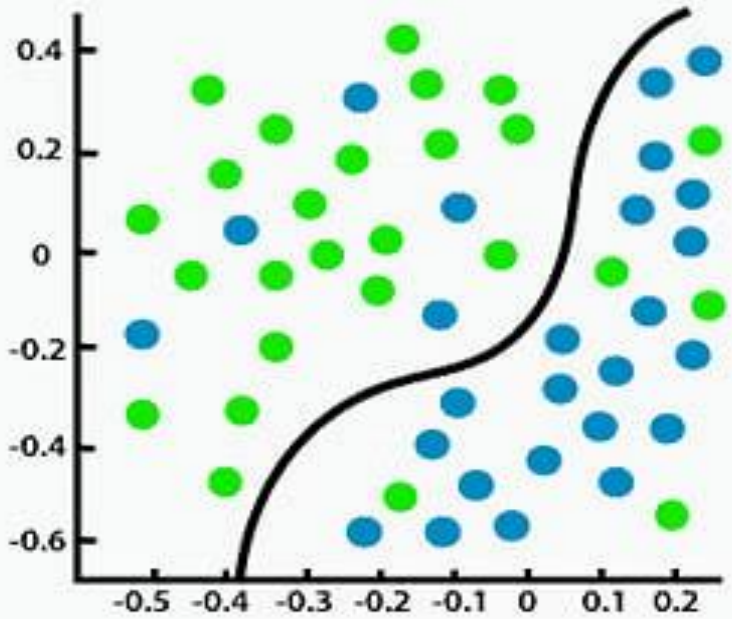
Stages of knowledge extraction





Regression

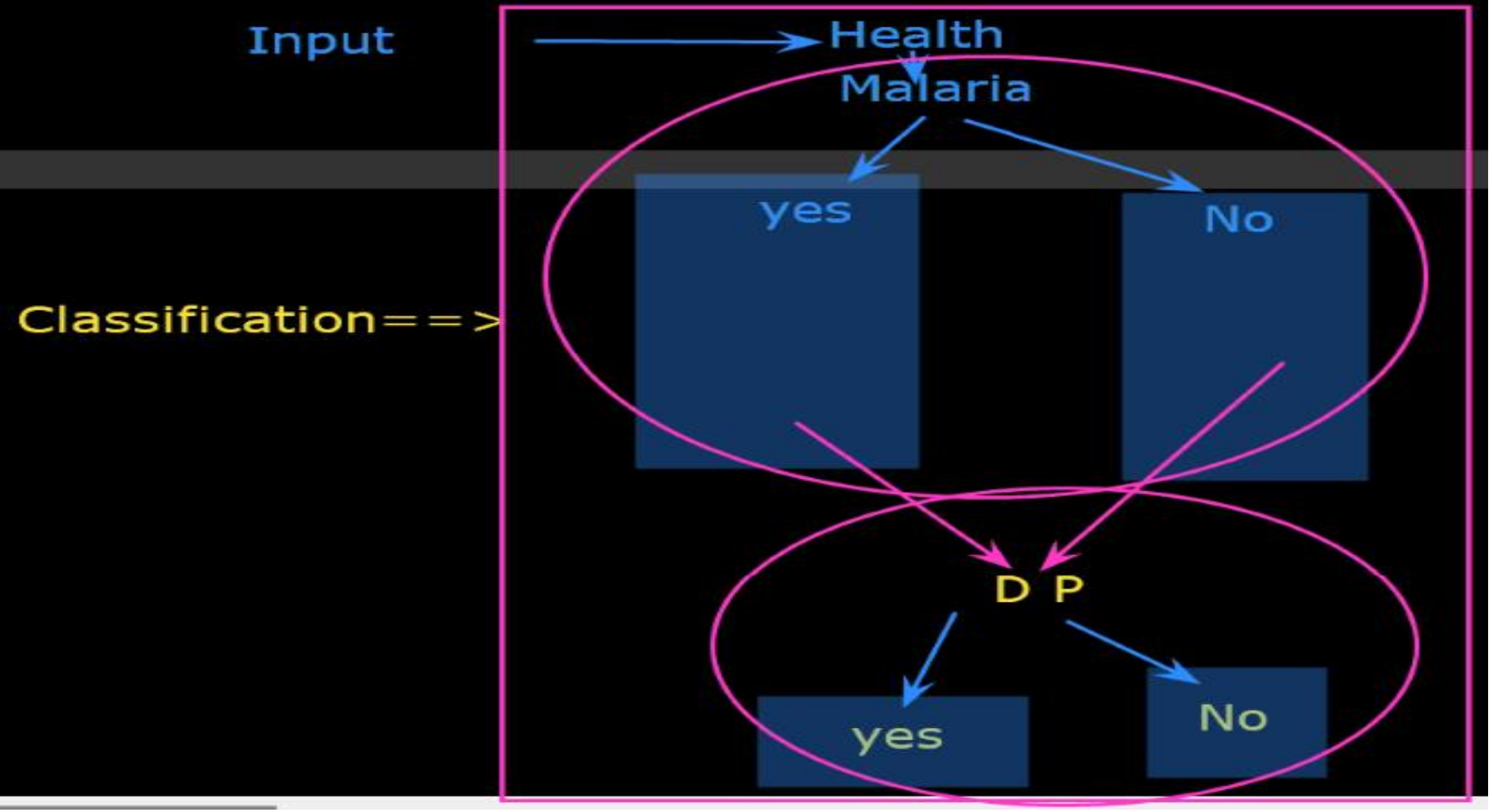
versus



Classification

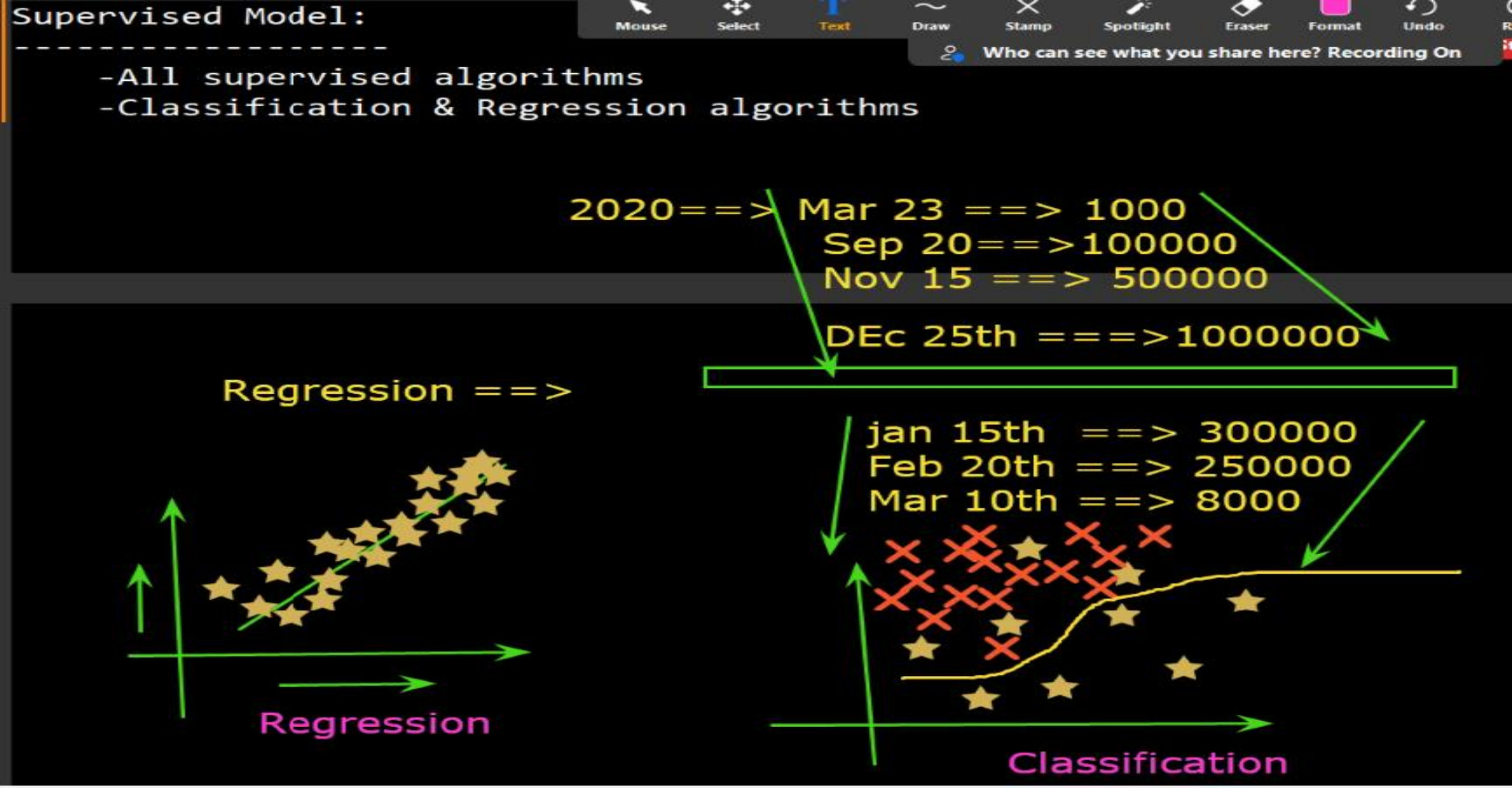
Supervised Model:

-All supervised algorithms
-Classification & Regression algorithms



Supervised Model:

-All supervised algorithms
-Classification & Regression algorithms

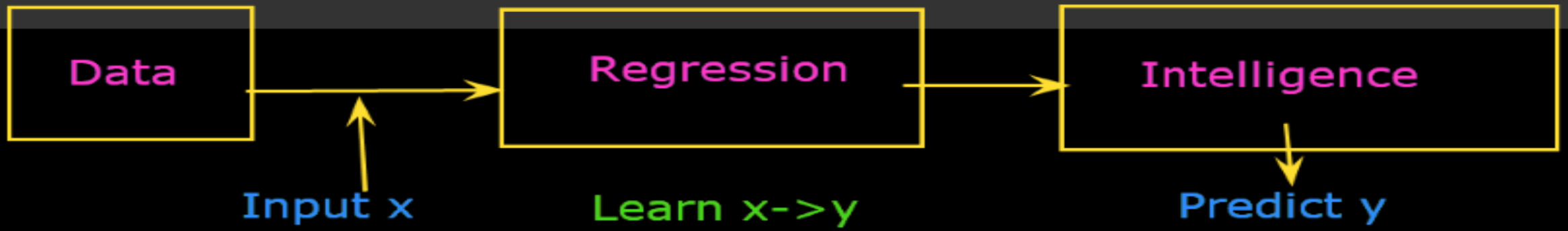


Regression:

- prediction of continuous features
- prediction of a features
- relation between features
- multi-variate regression

Mouse Select Text Draw Stamp Spotlight Eraser Format Undo

Who can see what you share here? Recording On



Regression:

- prediction of continuous features
- prediction of a features
- relation between features
- multi-variate regression

x=2, y=4
x=3, y=9
x=4, y=16

Input x

Learn x->y

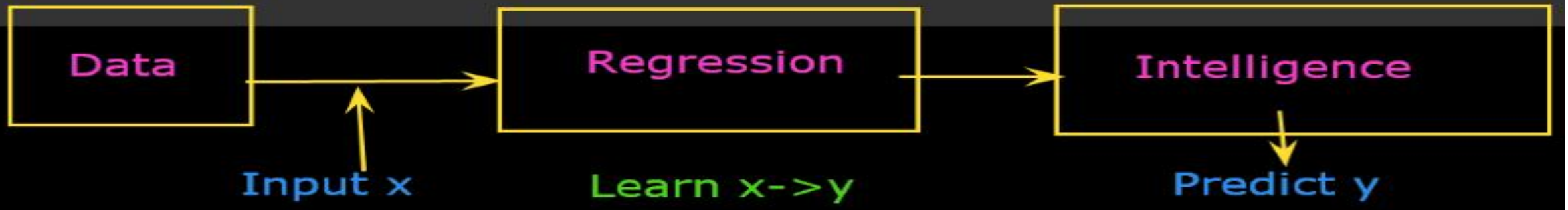
Predict y

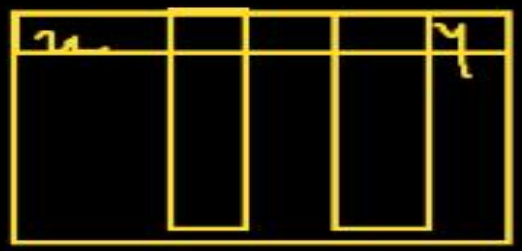
Regression:

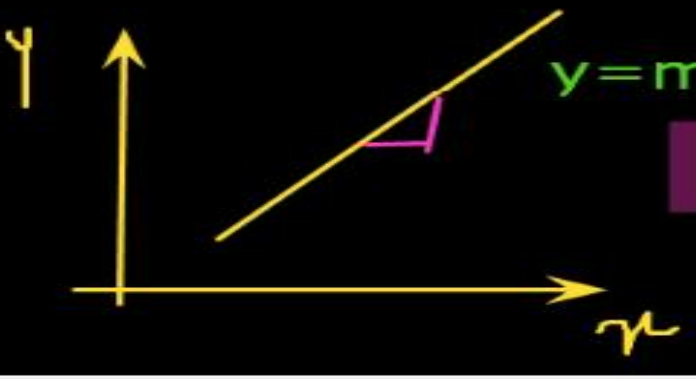
-prediction of continuous features
-prediction of a features
-relation between features
-multi-variate regression

$y=ax1+a2x2+a3x3+c$

x=2, y=4
x=3, y=9
x=4, y=16







y=2x+5

x=1, y=7
x=2, y=9

jupyter Day4 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code Commit

In [58]:
```
dataset= pd.read_csv('D:Data.csv')
dataset
```

Out[58]:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

Training Model

Testing Model