NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Uni-Variate Linear Regression

Sourav SEN GUPTA
Lecturer, SCSE, NTU

Data Science

Uni-Variate Regression

Machine Learning

Are variables mutually dependent?

How to find relation between them?

How to predict one using another?

How to optimally learn from the Data?

Data Science

The Pokemon Dataset

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
430	Honchkrow	Dark	Flying	505	100	125	52	105	52	71	4	False
338	Solrock	Rock	Psychic	440	70	95	85	55	65	70	3	False
32	Nidoran♂	Poison	NaN	273	46	57	40	40	40	50	1	False
442	Spiritomb	Ghost	Dark	485	50	92	108	92	108	35	4	False
480	Uxie	Psychic	NaN	580	75	75	130	75	130	95	4	True
536	Palpitoad	Water	Ground	384	75	65	55	65	55	69	5	False
360	Wynaut	Psychic	NaN	260	95	23	48	23	48	23	3	False
478	Froslass	Ice	Ghost	480	70	80	70	80	70	110	4	False
76	Golem	Rock	Ground	495	80	120	130	55	65	45	1	False
177	Natu	Psychic	Flying	320	40	50	45	70	45	70	2	False

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon

	HP
count	800.000000
mean	69.258750
std	25.534669
min	1.000000
25%	50.000000
50%	65.000000
75%	80.000000
max	255.000000

	Total
count	800.00000
mean	435.10250
std	119.96304
min	180.00000
25%	330.00000
50%	450.00000
75%	515.00000
max	780.00000

Data Science

Bi-Variate Exploration

Statistical Summary

HP Hit Points of a Pokemon

Total Total Points of a Pokemon

Machine Learning Questions

- What is the mutual relationship?
- Can we predict Total given HP?

	HP
count	600.000000
mean	69.026667
std	26.611267
min	1.000000
25%	50.000000
50%	65.000000
75%	80.000000
max	255.000000

	Total
count	600.000000
mean	432.715000
std	122.365283
min	180.000000
25%	325.000000
50%	440.000000
75%	515.000000
max	780.000000

Train Dataset

75% of the Data

Data Science

Bi-Variate Exploration

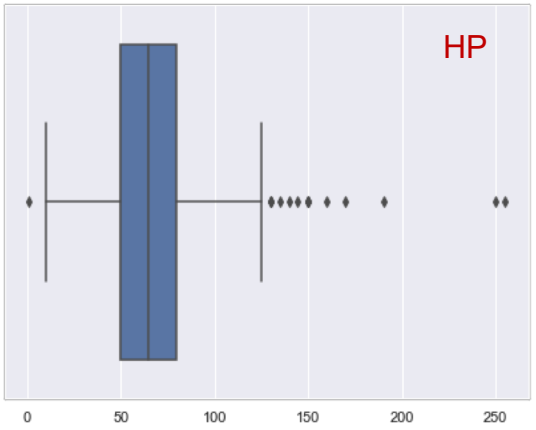
Split the Dataset

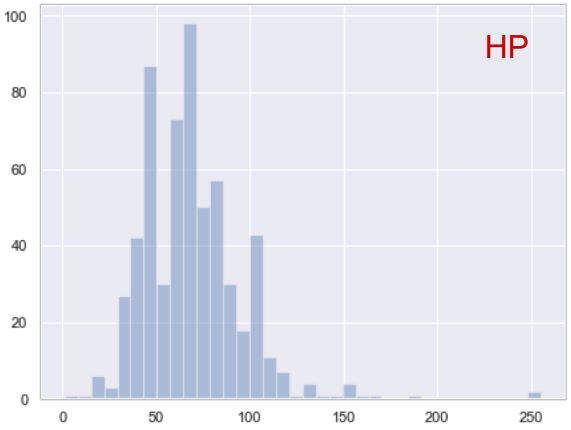
Train Used to train the model

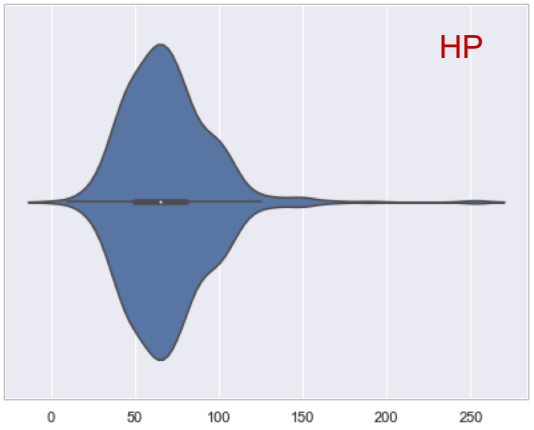
Test Used to test the model

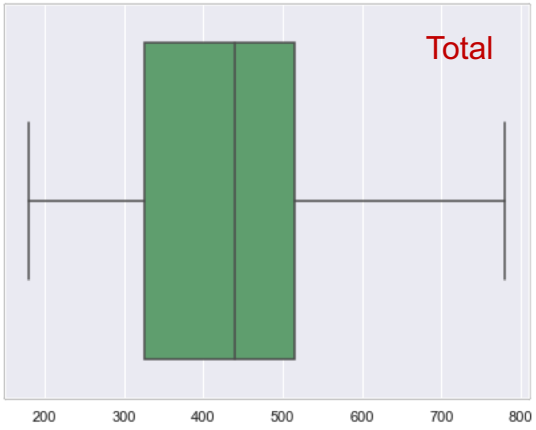
The Objectives

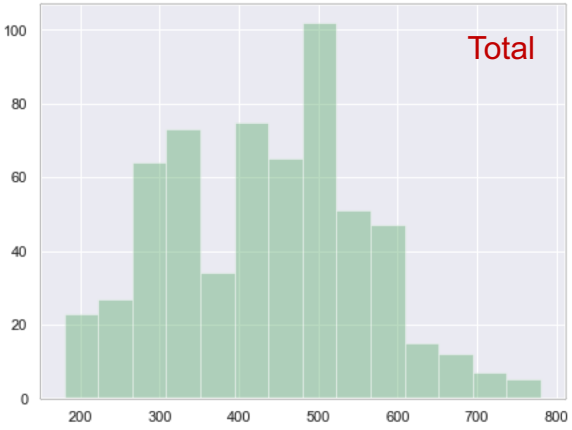
- Learn the relationship from Train
- Try to predict the Total on Test

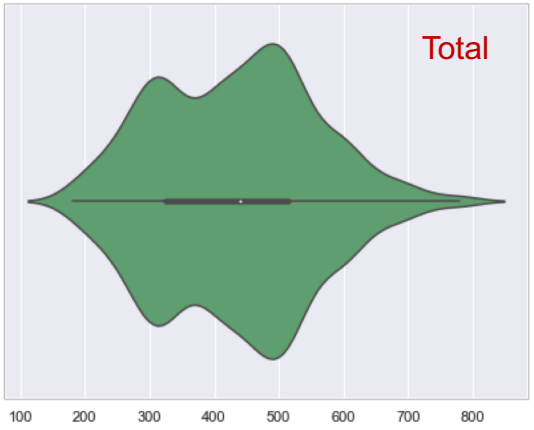

HP
0
50
100
150
200
250


HP
100
80
60
40
20
0
0
50
100
150
200
250


HP
0
50
100
150
200
250


Total
200
300
400
500
600
700
800


Total
100
80
60
40
20
0
200
300
400
500
600
700
800


Total
100
200
300
400
500
600
700
800

Data Science

Bi-Variate Exploration

Correlation Matrix and Plot

	Total	HP
Total	1.000000	0.590826
HP	0.590826	1.000000

Natural Intuition

Dependence of HP and Attack

Statistical Intuition

No Dependence **Corr = 0**

Perfect Positive **Corr = + 1**

Perfect Negative **Corr = - 1**

Data Science

Bi-Variate Exploration

Statistical Relation

Total	Plotted along Y axis
HP	Plotted along X axis

The Objectives

- Learn the relationship from Train
- Try to predict the Total on Test

Data Science

Uni-Variate Regression

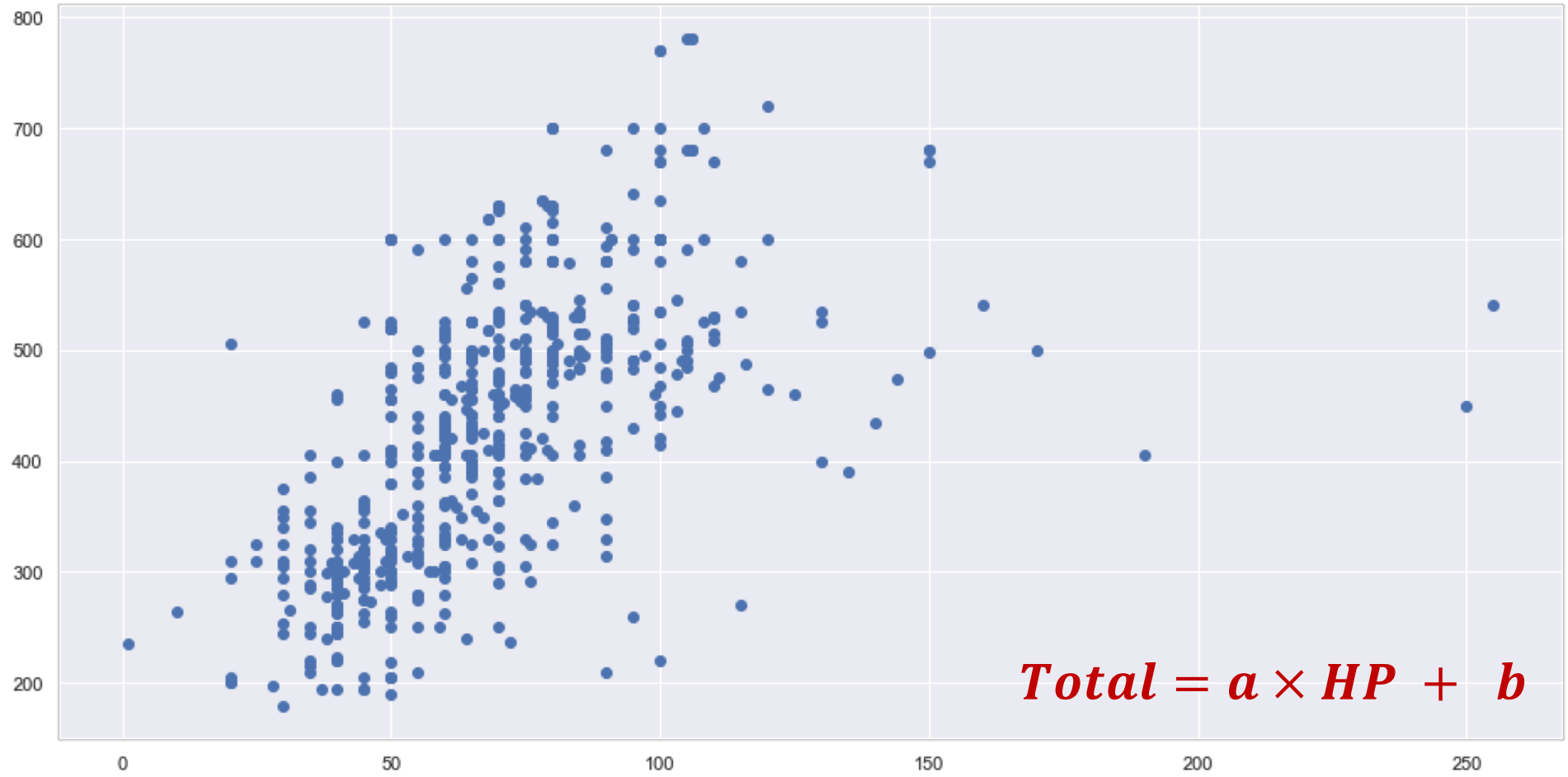
Statistical Modeling

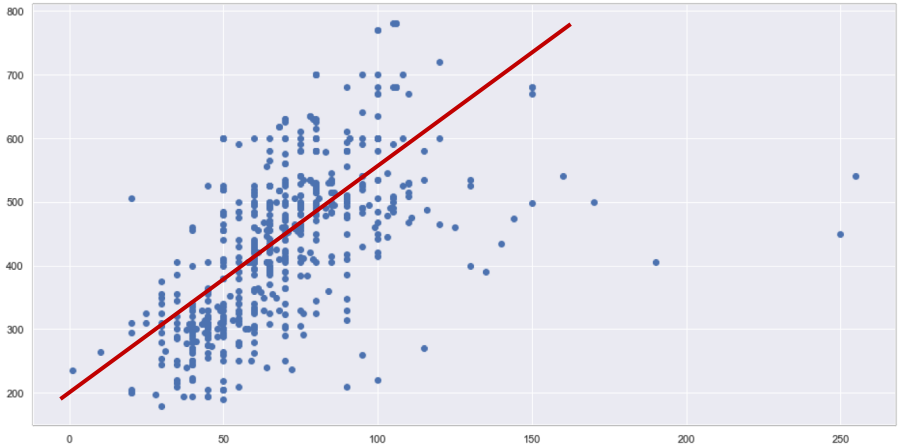
Hypothesize a Linear Model

$$Total = a \times HP + b + \epsilon$$

The Objectives

- Learn the parameters of the Model
- Try to use the Model for Prediction


800
700
600
500
400
300
200
0
50
100
150
200
250
$Total = a \times HP + b$



Data Science

Uni-Variate Regression

Algorithmic Optimization

Hypothesize a Linear Model

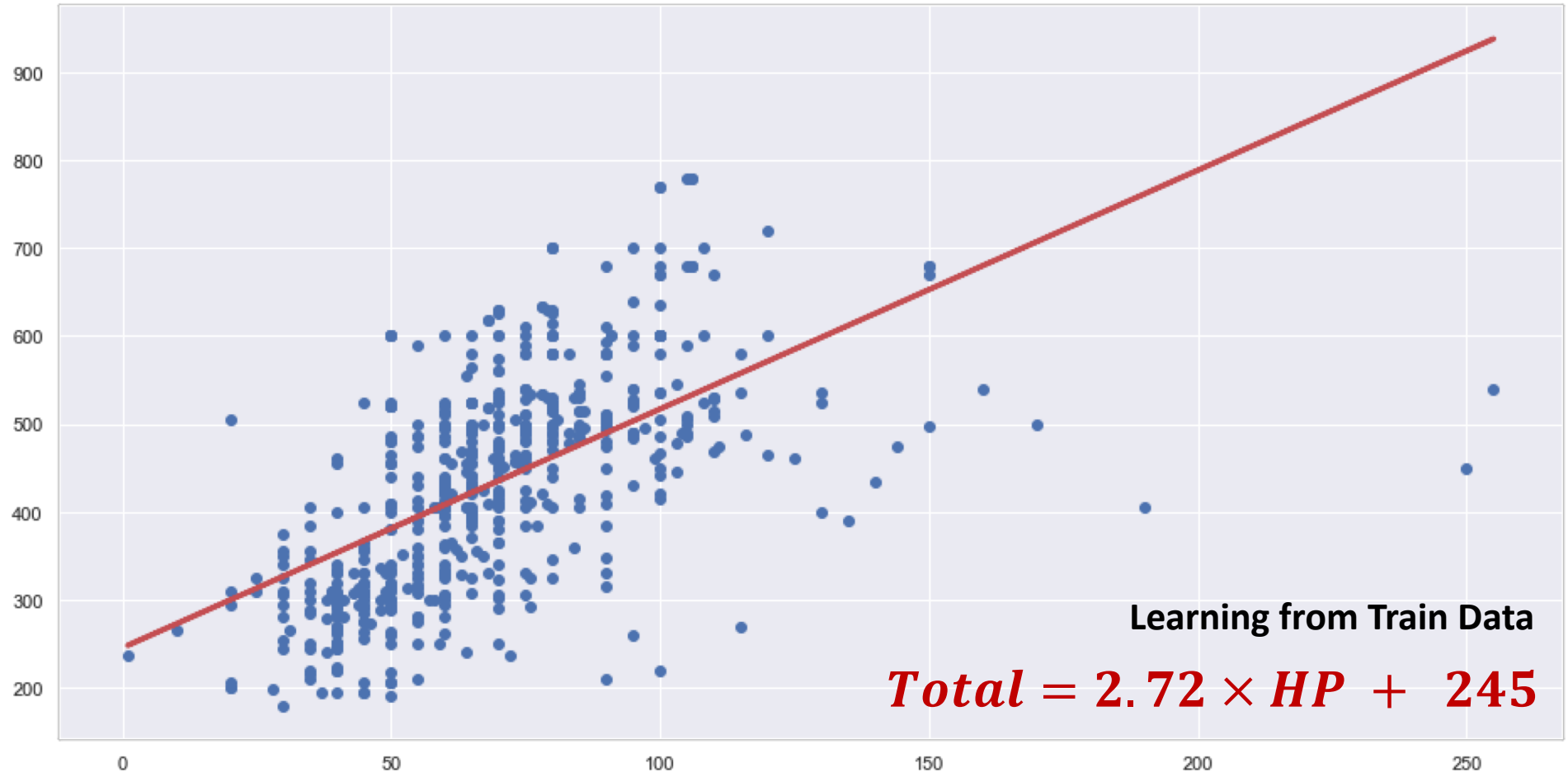
$$\boldsymbol{Total = a \times HP + b + \epsilon}$$

Steps in Linear Regression

Guess parameters ***a*** and ***b*** in the model
Predict the values of ***Total*** in Train Data
Calculate the ***Errors*** compared to actual
Tune the parameters to minimize ***Errors***

Cost Function to Minimize

$$J(a,b) = \sum (Total - a \times HP - b)^2$$


900
800
700
600
500
400
300
200
0
50
100
150
200
250
Learning from Train Data
$Total = 2.72 \times HP + 245$


700
600
500
400
300
200
40
60
80
100
120
140
160
Prediction on Test Data
$Total = 2.72 \times HP + 245$

Data Science

Uni-Variate Regression

Goodness of Fit of the Model

Hypothesis : The Linear Model

$$\boldsymbol{Total = a \times HP + b + \epsilon}$$

Explained Variance (R^2)

$$R^2 = 1 - \frac{\sum(Total - a \times HP - b)^2}{\sum(Total - \overline{Total})^2}$$

The higher the R^2 the better the Model

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum(Total - a \times HP - b)^2$$

The lower the MSE the better the Model

Data Science Pipeline

Machine Learning

How to learn from the acquired Data?

How to model the acquired Data?

How to predict on new Data?

How to optimally learn from the Data?