NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Data Science

Linear Regression

Sourav SEN GUPTA, Lecturer

School of Computer Science and Engg.
Nanyang Technological University

First Motivation for Data Science

PREDICTION IN BUSINESS

The Science

Common Problems

Prediction : Numeric

How Much?
How Many?

What is the expected Sales of a new SingTel store at North Spine?

Is it profitable to open?

https://www.youtube.com/watch?v=0XyV91VYrDs

The Science

Common Solutions

Prediction : Numeric

Regression

Prediction : Categorical

Classification

Detection : Anomalies

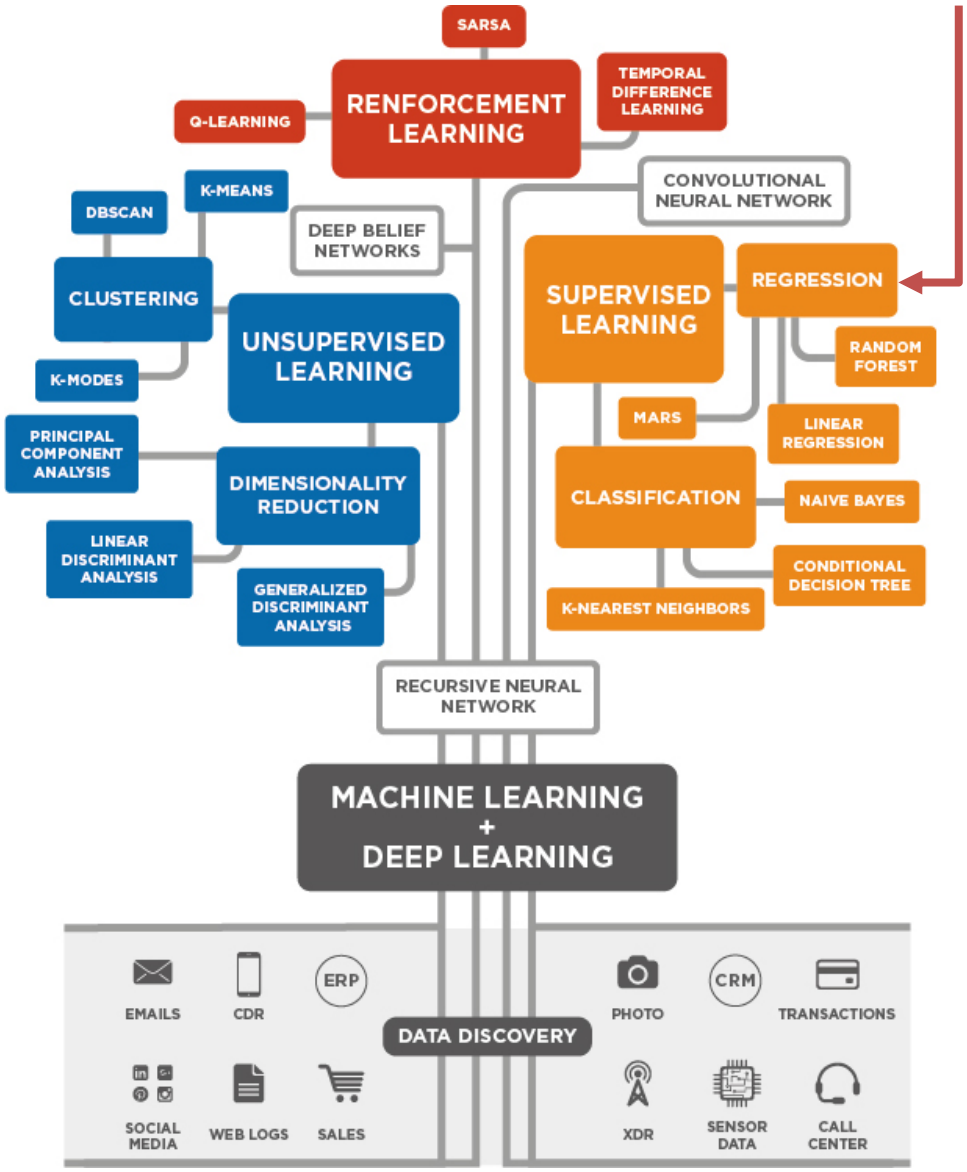
Anomaly Detection

Detection : Structure

Clustering and Dim-Red

Decision for Action

Reinforcement Learning



http://blogs.teradata.com/data-points/tree-machine-learning-algorithms/

What is the expected Sales at the *new* Store?

WHAT DATA DO YOU WANT?

The “Sales” Data

200 *similar* stores of the Company

22.1	10.4	9.3	18.5	12.9	7.2	11.8	13.2	4.8	10.6	8.6	17.4	9.2	9.7	19.0	22.4	12.5	24.4	11.3	14.6
18.0	12.5	5.6	15.5	9.7	12.0	15.0	15.9	18.9	10.5	21.4	11.9	9.6	17.4	9.5	12.8	25.4	14.7	10.1	21.5
16.6	17.1	20.7	12.9	8.5	14.9	10.6	23.2	14.8	9.7	11.4	10.7	22.6	21.2	20.2	23.7	5.5	13.2	23.8	18.4
8.1	24.2	15.7	14.0	18.0	9.3	9.5	13.4	18.9	22.3	18.3	12.4	8.8	11.0	17.0	8.7	6.9	14.2	5.3	11.0
11.8	12.3	11.3	13.6	21.7	15.2	12.0	16.0	12.9	16.7	11.2	7.3	19.4	22.2	11.5	16.9	11.7	15.5	25.4	17.2
11.7	23.8	14.8	14.7	20.7	19.2	7.2	8.7	5.3	19.8	13.4	21.8	14.1	15.9	14.6	12.6	12.2	9.4	15.9	6.6
15.5	7.0	11.6	15.2	19.7	10.6	6.6	8.8	24.7	9.7	1.6	12.7	5.7	19.6	10.8	11.6	9.5	20.8	9.6	20.7
10.9	19.2	20.1	10.4	11.4	10.3	13.2	25.4	10.9	10.1	16.1	11.6	16.6	19.0	15.6	3.2	15.3	10.1	7.3	12.9
14.4	13.3	14.9	18.0	11.9	11.9	8.0	12.2	17.1	15.0	8.4	14.5	7.6	11.7	11.5	27.0	20.2	11.7	11.8	12.6
10.5	12.2	8.7	26.2	17.6	22.6	10.3	17.3	15.9	6.7	10.8	9.9	5.9	19.6	17.3	7.6	9.7	12.8	25.5	13.4

Can you predict the estimated Sales of the *new* store?

The “Sales” Data

200 *similar* stores of the Company

22.1	10.4	9.3	18.5	12.9	7.2	11.8	13.2	4.8	10.6	8.6	17.4	9.2	9.7	19.0	22.4	12.5	24.4	11.3	14.6
18.0	12.5	5.6	15.5	9.7	12.0	15.0	15.9	18.9	10.5	21.4	11.9	9.6	17.4	9.5	12.8	25.4	14.7	10.1	21.5
16.6	17.1	20.7	12.9	8.5	14.9	10.6	23.2	14.8	9.7	11.4	10.7	22.6	21.2	20.2	23.7	5.5	13.2	23.8	18.4
8.1	24.2	15.7	14.0	18.0	9.3	9.5	13.4	18.9	22.3	18.3	12.4	8.8	11.0	17.0	8.7	6.9	14.2	5.3	11.0
11.8	12.3	11.3	13.6	21.7	15.2	12.0	16.0	12.9	16.7	11.2	7.3	19.4	22.2	11.5	16.9	11.7	15.5	25.4	17.2
11.7	23.8	14.8	14.7	20.7	19.2	7.2	8.7	5.3	19.8	13.4	21.8	14.1	15.9	14.6	12.6	12.2	9.4	15.9	6.6
15.5	7.0	11.6	15.2	19.7	10.6	6.6	8.8	24.7	9.7	1.6	12.7	5.7	19.6	10.8	11.6	9.5	20.8	9.6	20.7
10.9	19.2	20.1	10.4	11.4	10.3	13.2	25.4	10.9	10.1	16.1	11.6	16.6	19.0	15.6	3.2	15.3	10.1	7.3	12.9
14.4	13.3	14.9	18.0	11.9	11.9	8.0	12.2	17.1	15.0	8.4	14.5	7.6	11.7	11.5	27.0	20.2	11.7	11.8	12.6
10.5	12.2	8.7	26.2	17.6	22.6	10.3	17.3	15.9	6.7	10.8	9.9	5.9	19.6	17.3	7.6	9.7	12.8	25.5	13.4

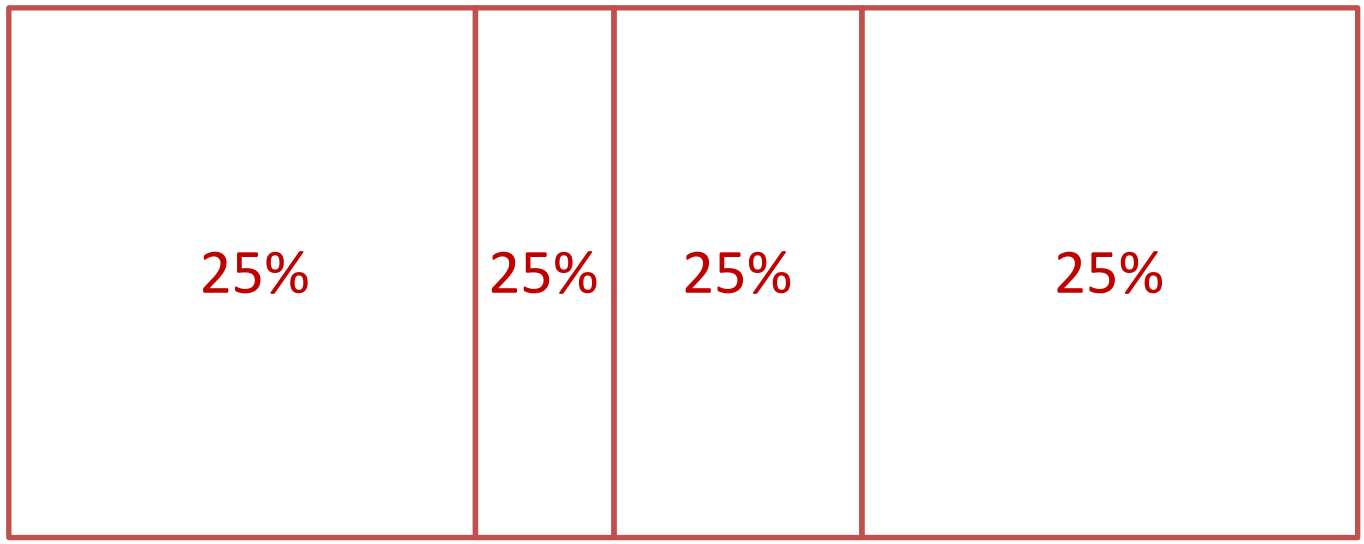
On an *average*, Sales is 14.0225 (Mean)

The *estimate* is wrong by 5.204 (Standard Deviation)

The “Sales” Data

200 *similar* stores of the Company

Summary Statistics	Min.	Q1	Median	Q3	Max.
	1.60	10.38	12.90	17.40	27.00



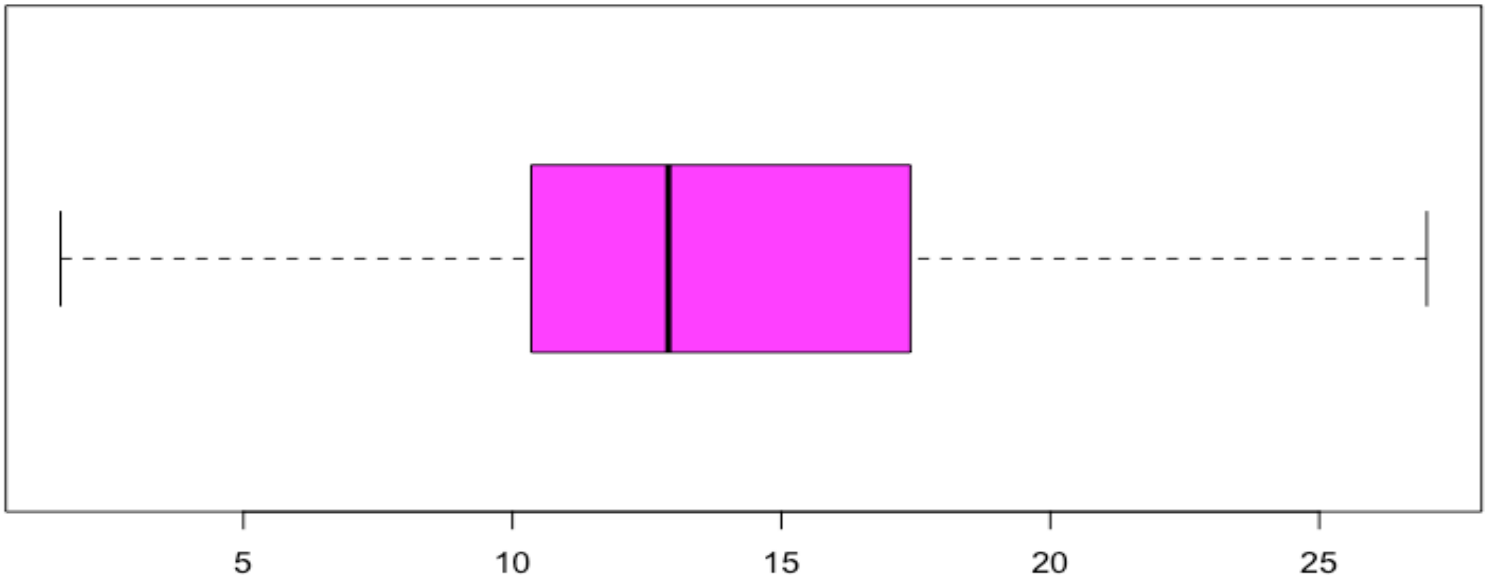
The “Sales” Data

200 *similar* stores of the Company

Summary Statistics

Min.	Q1	Median	Q3	Max.
1.60	10.38	12.90	17.40	27.00

Box-Plot



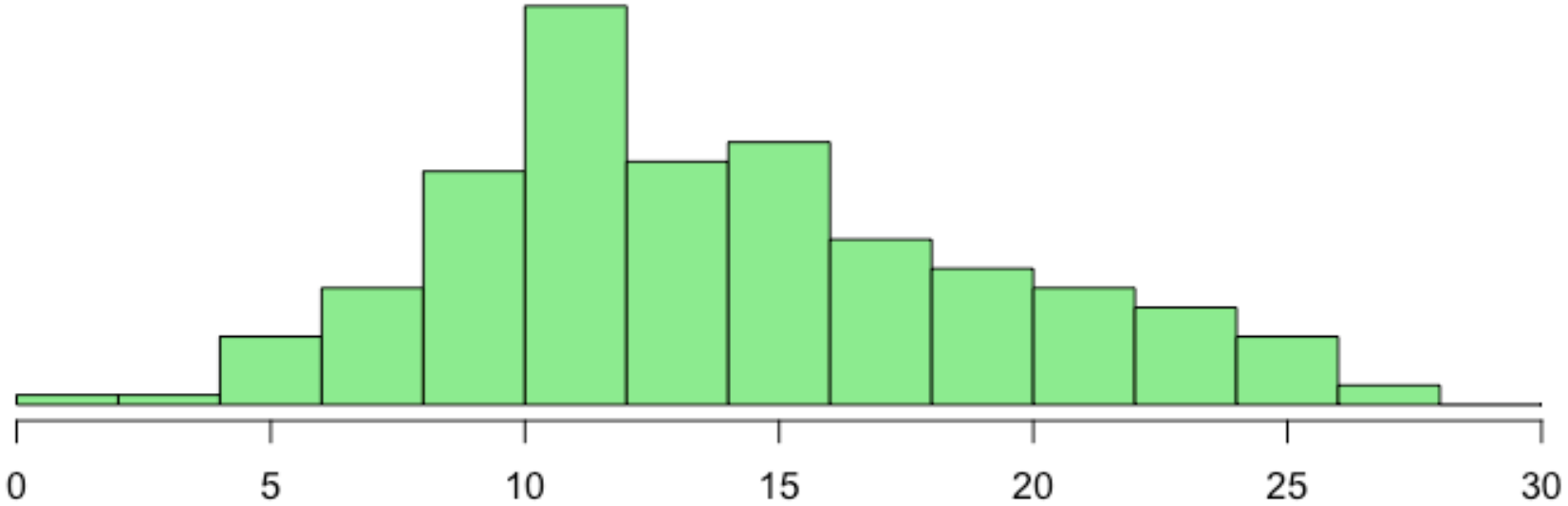
The “Sales” Data

200 *similar* stores of the Company

Summary Statistics

Min.	Q1	Median	Q3	Max.
1.60	10.38	12.90	17.40	27.00

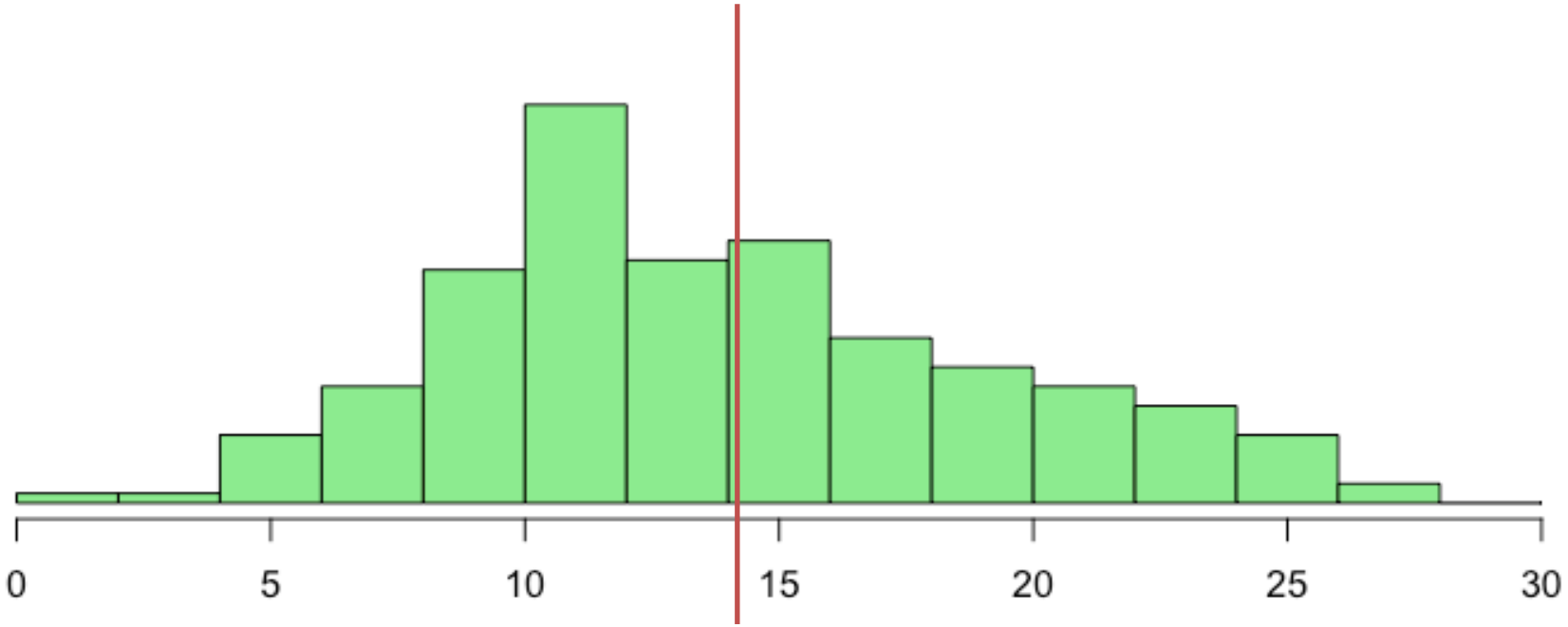
Histogram



Box-Plot

The “Sales” Data

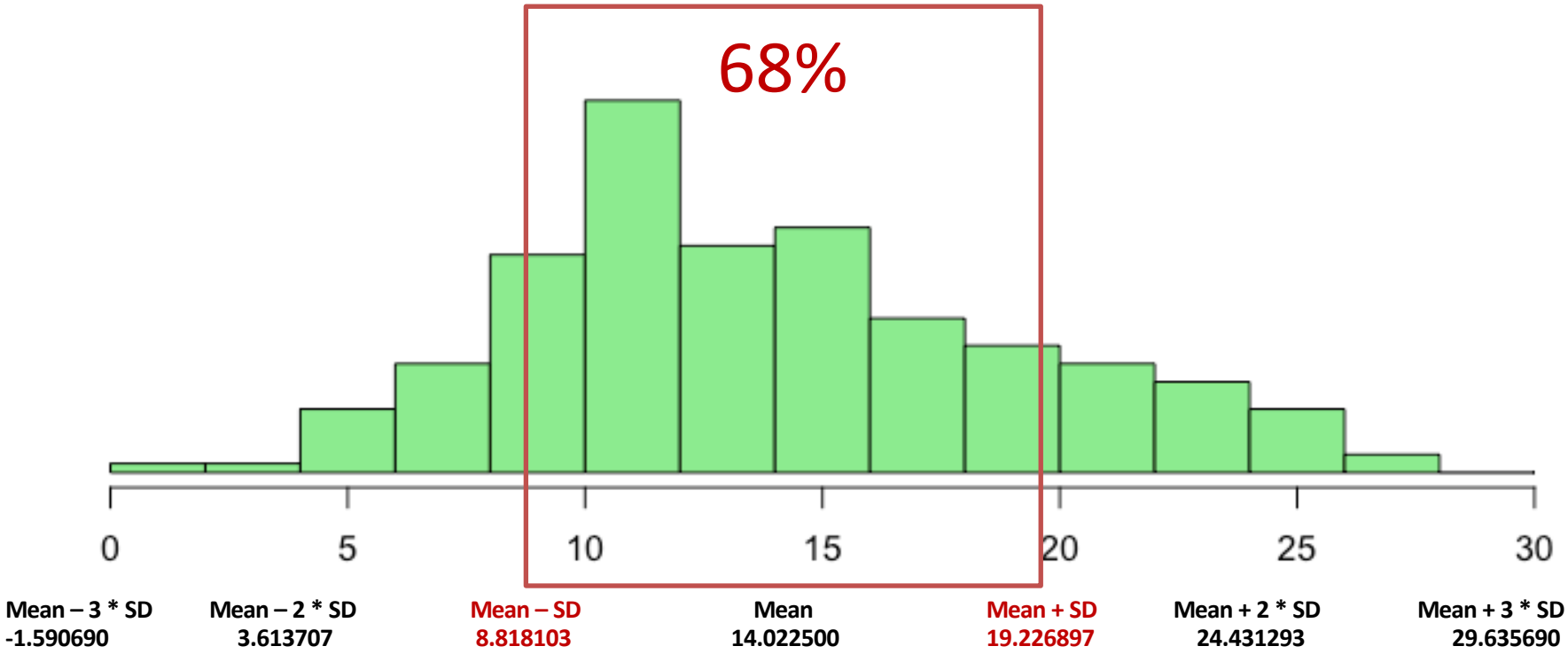
200 *similar* stores of the Company



Mean – 3 * SD	Mean – 2 * SD	Mean – SD	Mean	Mean + SD	Mean + 2 * SD	Mean + 3 * SD
-1.590690	3.613707	8.818103	14.022500	19.226897	24.431293	29.635690

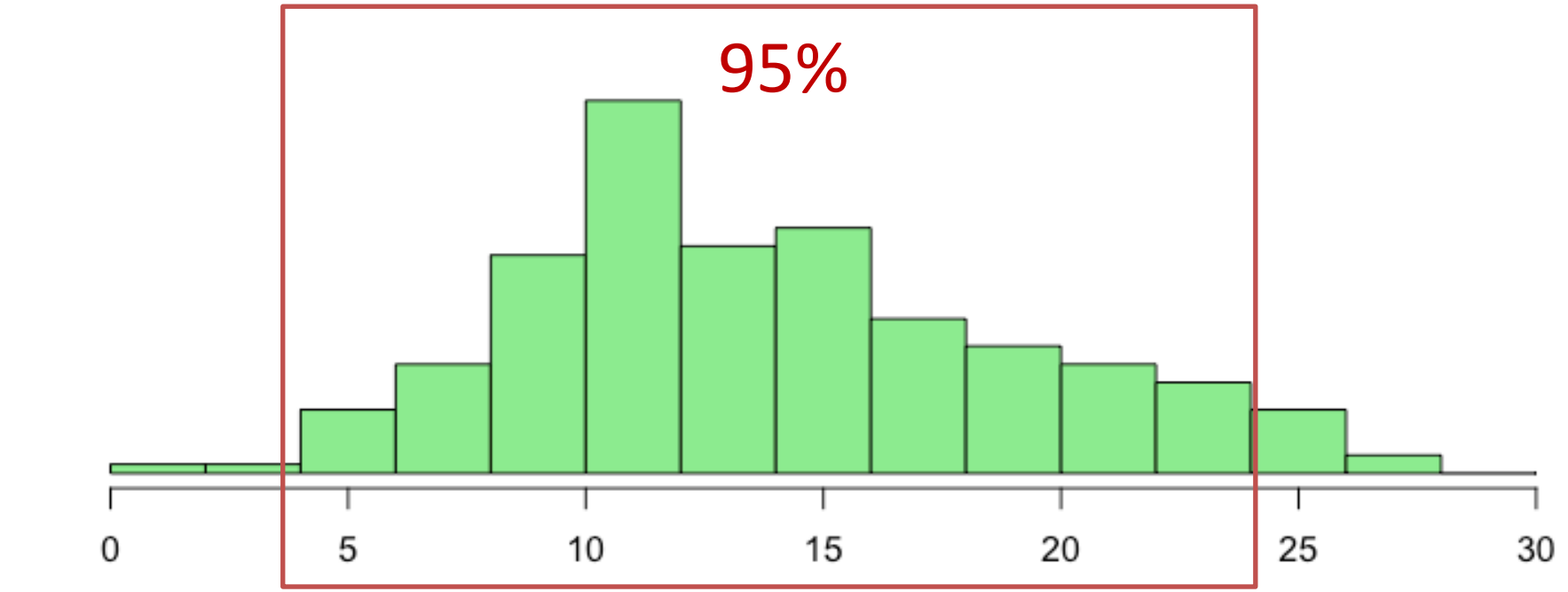
The “Sales” Data

200 *similar* stores of the Company



The “Sales” Data

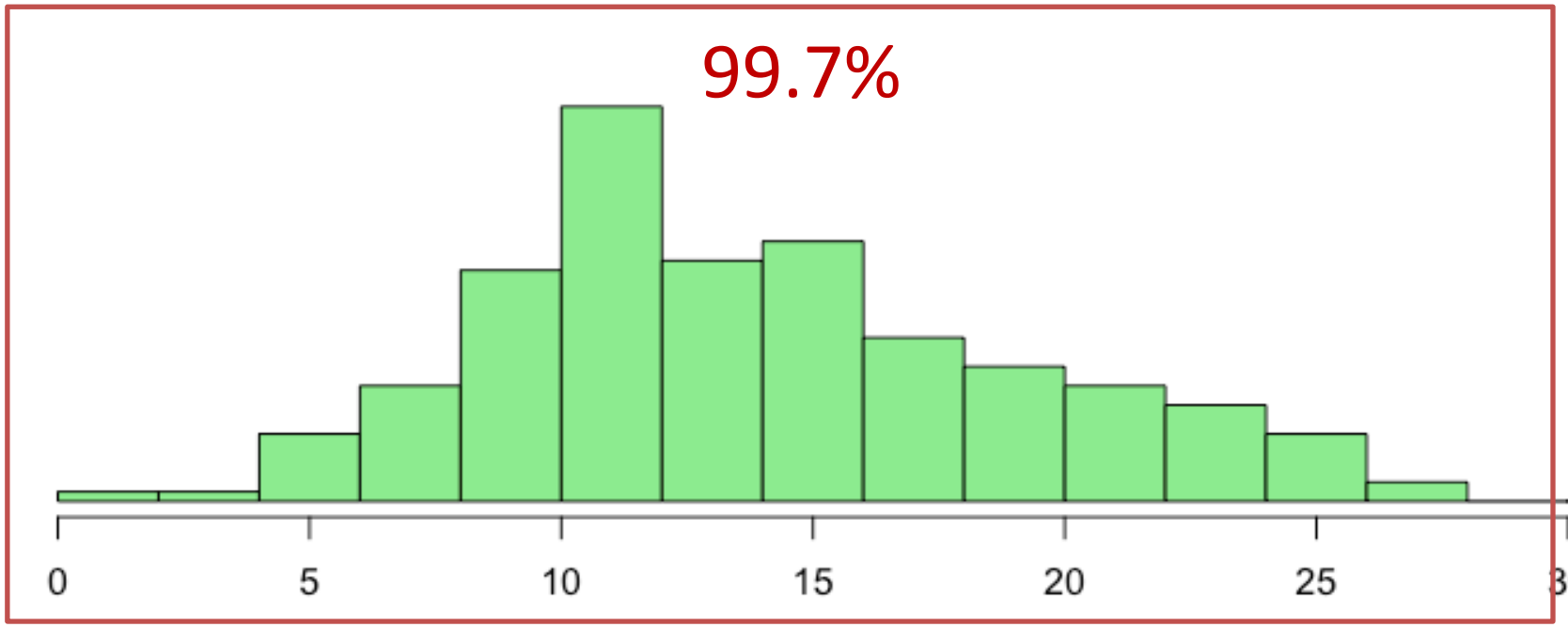
200 *similar* stores of the Company



Mean – 3 * SD	Mean – 2 * SD	Mean – SD	Mean	Mean + SD	Mean + 2 * SD	Mean + 3 * SD
-1.590690	3.613707	8.818103	14.022500	19.226897	24.431293	29.635690

The “Sales” Data

200 *similar* stores of the Company



Mean – 3 * SD	Mean – 2 * SD	Mean – SD	Mean	Mean + SD	Mean + 2 * SD	Mean + 3 * SD
-1.590690	3.613707	8.818103	14.022500	19.226897	24.431293	29.635690

The “Sales” Data

200 *similar* stores of the Company

22.1	10.4	9.3	18.5	12.9	7.2	11.8	13.2	4.8	10.6	8.6	17.4	9.2	9.7	19.0	22.4	12.5	24.4	11.3	14.6
18.0	12.5	5.6	15.5	9.7	12.0	15.0	15.9	18.9	10.5	21.4	11.9	9.6	17.4	9.5	12.8	25.4	14.7	10.1	21.5
16.6	17.1	20.7	12.9	8.5	14.9	10.6	23.2	14.8	9.7	11.4	10.7	22.6	21.2	20.2	23.7	5.5	13.2	23.8	18.4
8.1	24.2	15.7	14.0	18.0	9.3	9.5	13.4	18.9	22.3	18.3	12.4	8.8	11.0	17.0	8.7	6.9	14.2	5.3	11.0
11.8	12.3	11.3	13.6	21.7	15.2	12.0	16.0	12.9	16.7	11.2	7.3	19.4	22.2	11.5	16.9	11.7	15.5	25.4	17.2
11.7	23.8	14.8	14.7	20.7	19.2	7.2	8.7	5.3	19.8	13.4	21.8	14.1	15.9	14.6	12.6	12.2	9.4	15.9	6.6
15.5	7.0	11.6	15.2	19.7	10.6	6.6	8.8	24.7	9.7	1.6	12.7	5.7	19.6	10.8	11.6	9.5	20.8	9.6	20.7
10.9	19.2	20.1	10.4	11.4	10.3	13.2	25.4	10.9	10.1	16.1	11.6	16.6	19.0	15.6	3.2	15.3	10.1	7.3	12.9
14.4	13.3	14.9	18.0	11.9	11.9	8.0	12.2	17.1	15.0	8.4	14.5	7.6	11.7	11.5	27.0	20.2	11.7	11.8	12.6
10.5	12.2	8.7	26.2	17.6	22.6	10.3	17.3	15.9	6.7	10.8	9.9	5.9	19.6	17.3	7.6	9.7	12.8	25.5	13.4

On an *average*, Sales is 14.0225 – this is the Mean value

Should be between 8.82 to 19.23 – with 68% *confidence*

The “Sales” Data

200 *similar* stores of the Company

22.1	10.4	9.3	18.5	12.9	7.2	11.8	13.2	4.8	10.6	8.6	17.4	9.2	9.7	19.0	22.4	12.5	24.4	11.3	14.6
18.0	12.5	5.6	15.5	9.7	12.0	15.0	15.9	18.9	10.5	21.4	11.9	9.6	17.4	9.5	12.8	25.4	14.7	10.1	21.5
16.6	17.1	20.7	12.9	8.5	14.9	10.6	23.2	14.8	9.7	11.4	10.7	22.6	21.2	20.2	23.7	5.5	13.2	23.8	18.4
8.1	24.2	15.7	14.0	18.0	9.3	9.5	13.4	18.9	22.3	18.3	12.4	8.8	11.0	17.0	8.7	6.9	14.2	5.3	11.0
11.8	12.3	11.3	13.6	21.7	15.2	12.0	16.0	12.9	16.7	11.2	7.3	19.4	22.2	11.5	16.9	11.7	15.5	25.4	17.2
11.7	23.8	14.8	14.7	20.7	19.2	7.2	8.7	5.3	19.8	13.4	21.8	14.1	15.9	14.6	12.6	12.2	9.4	15.9	6.6
15.5	7.0	11.6	15.2	19.7	10.6	6.6	8.8	24.7	9.7	1.6	12.7	5.7	19.6	10.8	11.6	9.5	20.8	9.6	20.7
10.9	19.2	20.1	10.4	11.4	10.3	13.2	25.4	10.9	10.1	16.1	11.6	16.6	19.0	15.6	3.2	15.3	10.1	7.3	12.9
14.4	13.3	14.9	18.0	11.9	11.9	8.0	12.2	17.1	15.0	8.4	14.5	7.6	11.7	11.5	27.0	20.2	11.7	11.8	12.6
10.5	12.2	8.7	26.2	17.6	22.6	10.3	17.3	15.9	6.7	10.8	9.9	5.9	19.6	17.3	7.6	9.7	12.8	25.5	13.4

On an *average*, Sales is 14.0225 – this is the Mean value

Should be between 3.61 to 24.43 – with 95% *confidence*

The “Sales” Data

200 *similar* stores of the Company

22.1	10.4	9.3	18.5	12.9	7.2	11.8	13.2	4.8	10.6	8.6	17.4	9.2	9.7	19.0	22.4	12.5	24.4	11.3	14.6
18.0	12.5	5.6	15.5	9.7	12.0	15.0	15.9	18.9	10.5	21.4	11.9	9.6	17.4	9.5	12.8	25.4	14.7	10.1	21.5
16.6	17.1	20.7	12.9	8.5	14.9	10.6	23.2	14.8	9.7	11.4	10.7	22.6	21.2	20.2	23.7	5.5	13.2	23.8	18.4
8.1	24.2	15.7	14.0	18.0	9.3	9.5	13.4	18.9	22.3	18.3	12.4	8.8	11.0	17.0	8.7	6.9	14.2	5.3	11.0
11.8	12.3	11.3	13.6	21.7	15.2	12.0	16.0	12.9	16.7	11.2	7.3	19.4	22.2	11.5	16.9	11.7	15.5	25.4	17.2
11.7	23.8	14.8	14.7	20.7	19.2	7.2	8.7	5.3	19.8	13.4	21.8	14.1	15.9	14.6	12.6	12.2	9.4	15.9	6.6
15.5	7.0	11.6	15.2	19.7	10.6	6.6	8.8	24.7	9.7	1.6	12.7	5.7	19.6	10.8	11.6	9.5	20.8	9.6	20.7
10.9	19.2	20.1	10.4	11.4	10.3	13.2	25.4	10.9	10.1	16.1	11.6	16.6	19.0	15.6	3.2	15.3	10.1	7.3	12.9
14.4	13.3	14.9	18.0	11.9	11.9	8.0	12.2	17.1	15.0	8.4	14.5	7.6	11.7	11.5	27.0	20.2	11.7	11.8	12.6
10.5	12.2	8.7	26.2	17.6	22.6	10.3	17.3	15.9	6.7	10.8	9.9	5.9	19.6	17.3	7.6	9.7	12.8	25.5	13.4

On an *average*, Sales is 14.0225 – this is the Mean value

Should be between -1.59 to 29.64 – with 99.7% *confidence*

Can you do any better than that …

IF I GIVE YOU MORE DATA?

Advertising Data

200 *similar* stores of the Company

X	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
⋮	⋮	⋮	⋮	⋮
196	38.2	3.7	13.8	7.6
197	94.2	4.9	8.1	9.7
198	177.0	9.3	6.4	12.8
199	283.6	42.0	66.2	25.5
200	232.1	8.6	8.7	13.4

How do you know if the extra data is at all **useful**?

Does advertisement has any **effect** on Sales?

Does advertisement has any **relation** with Sales?

Advertising Data

200 *similar* stores of the Company

TV	Radio	Newspaper	Sales
Min. : 0.70	Min. : 0.000	Min. : 0.30	Min. : 1.60
1st Qu.: 74.38	1st Qu.: 9.975	1st Qu.: 12.75	1st Qu.:10.38
Median :149.75	Median :22.900	Median : 25.75	Median :12.90
Mean :147.04	Mean :23.264	Mean : 30.55	Mean :14.02
3rd Qu.:218.82	3rd Qu.:36.525	3rd Qu.: 45.10	3rd Qu.:17.40
Max. :296.40	Max. :49.600	Max. :114.00	Max. :27.00

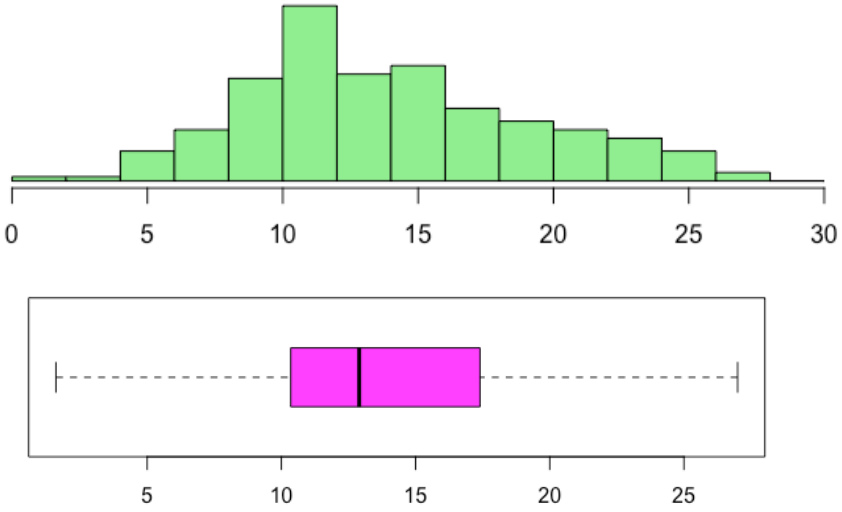
	TV	Radio	Newspaper	Sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
Radio	0.05480866	1.00000000	0.35410375	0.5762226
Newspaper	0.05664787	0.35410375	1.00000000	0.2282990
Sales	0.78222442	0.57622257	0.22829903	1.0000000

Mutual Correlations

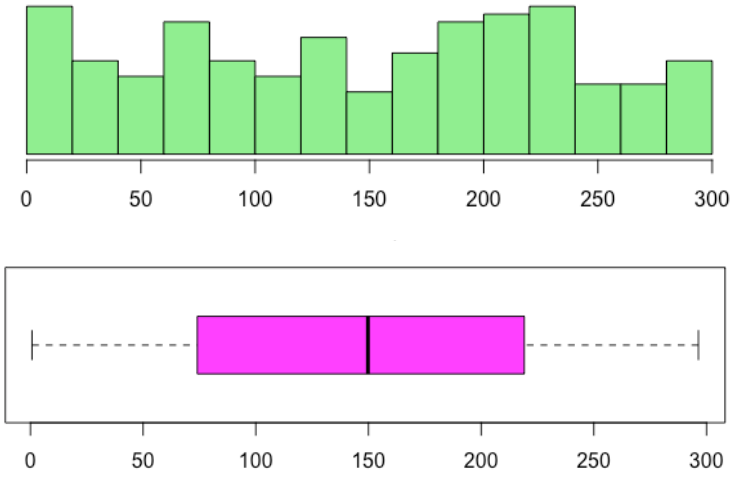
Sales vs TV advertising

200 *similar* stores of the Company

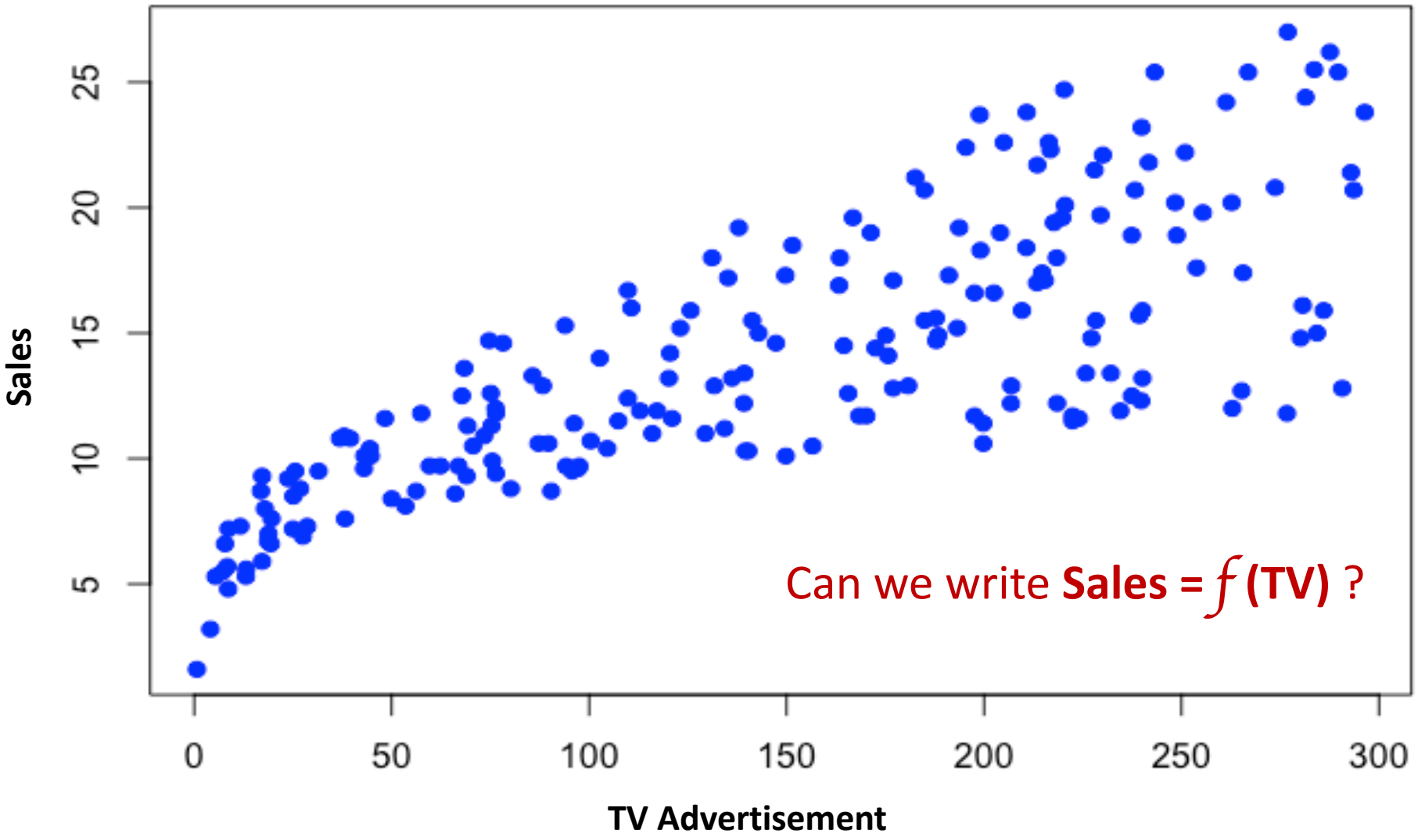
Distribution of Sales



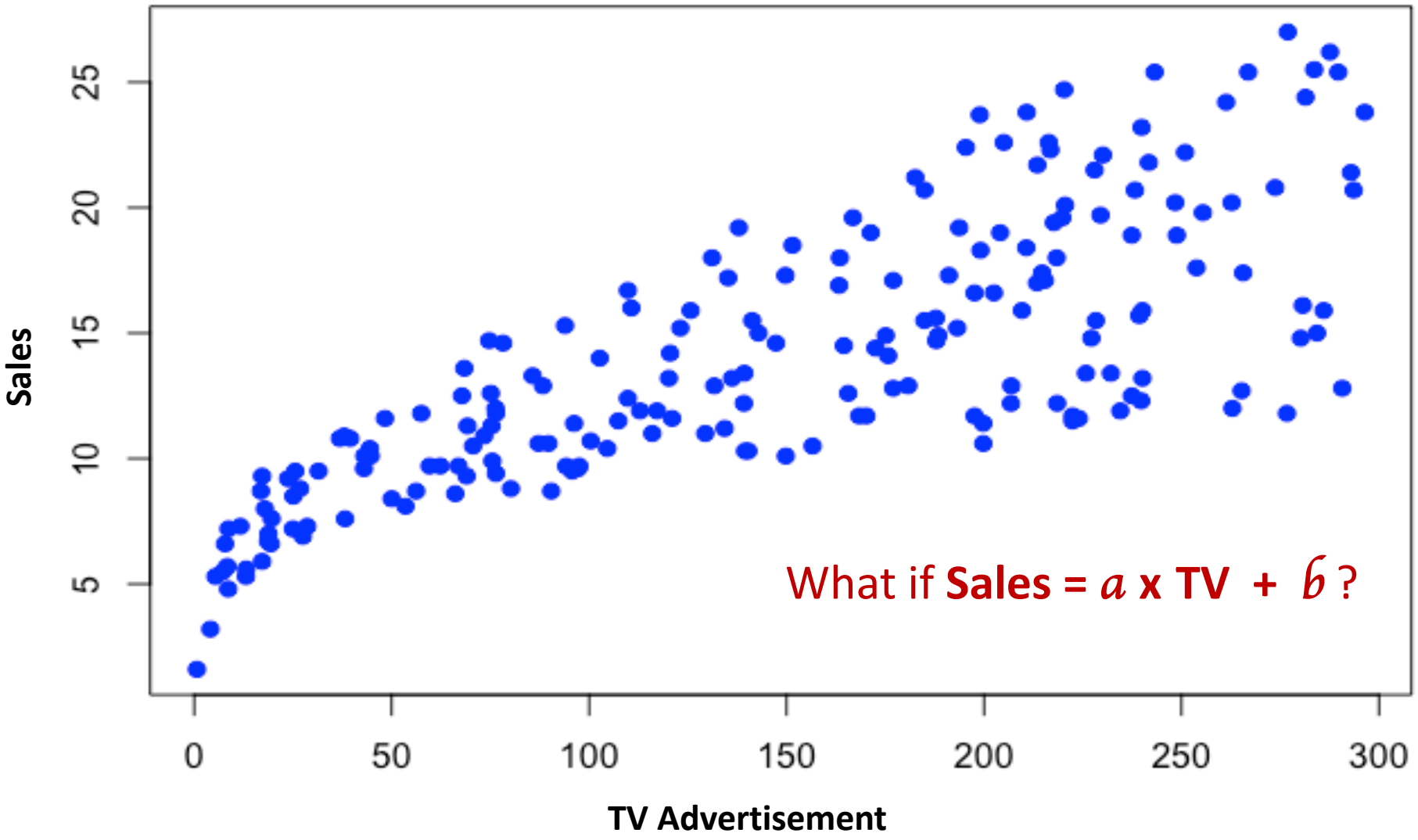
Distribution of TV adv.



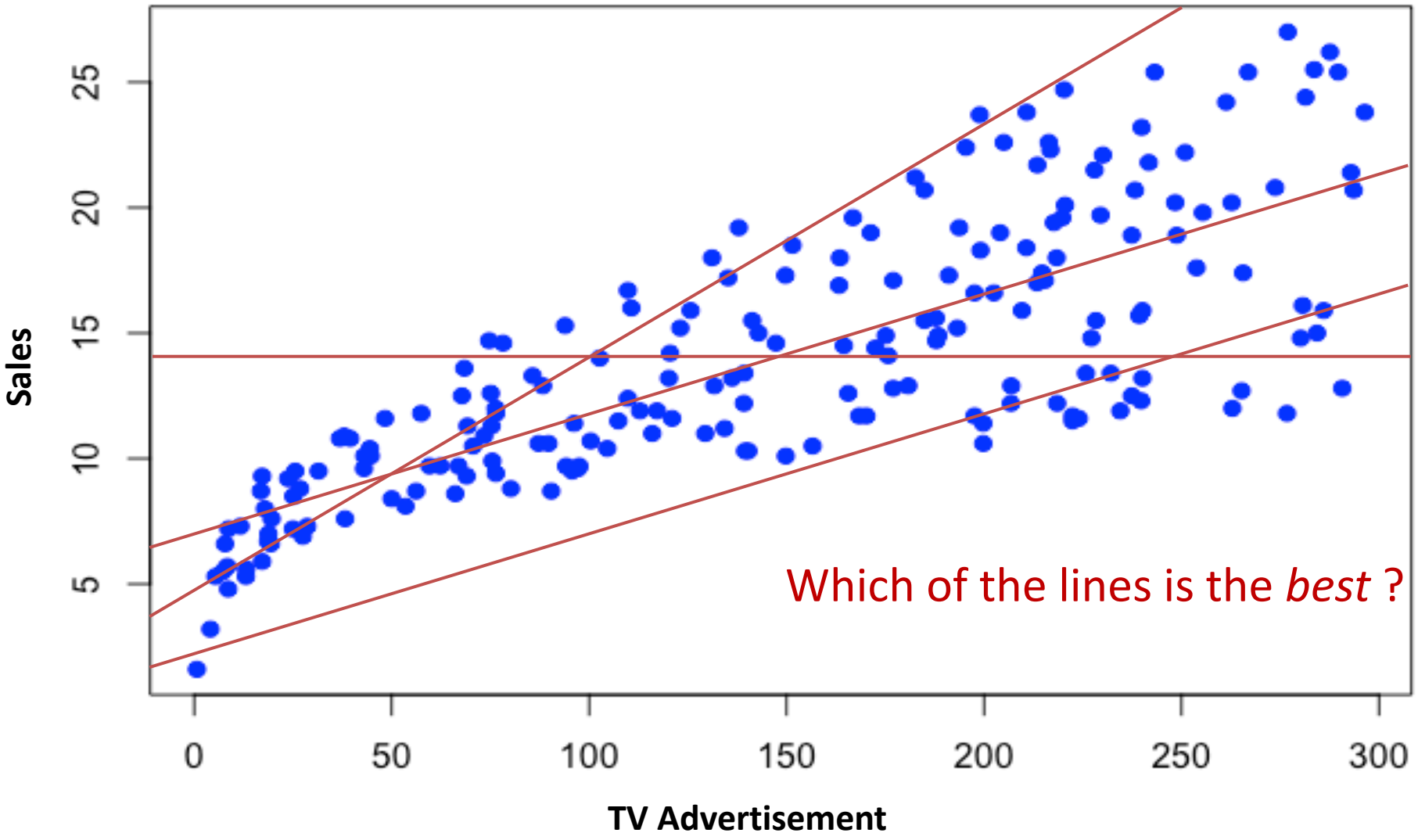
Distribution of Sales with respect to TV advertisement



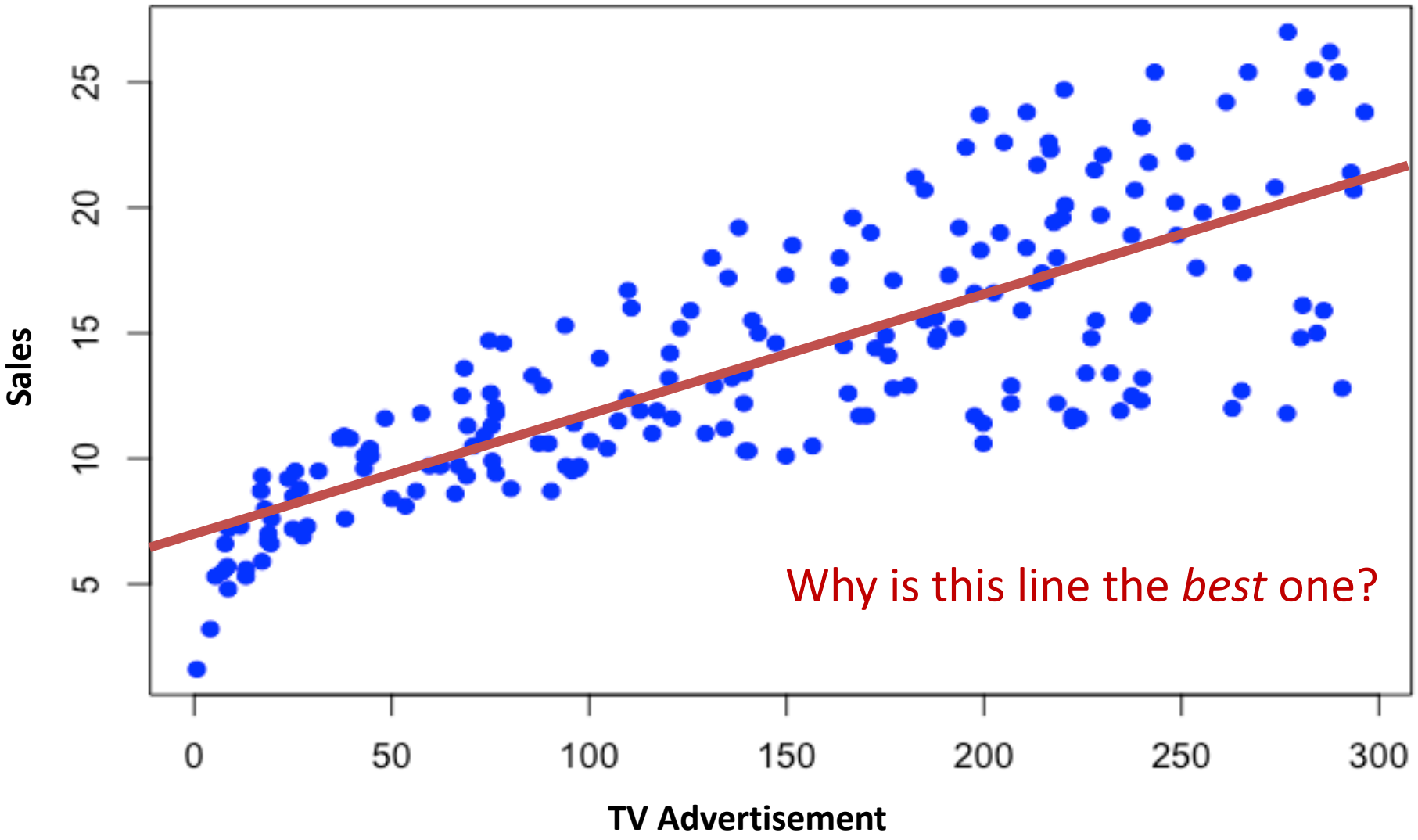
Distribution of Sales with respect to TV advertisement



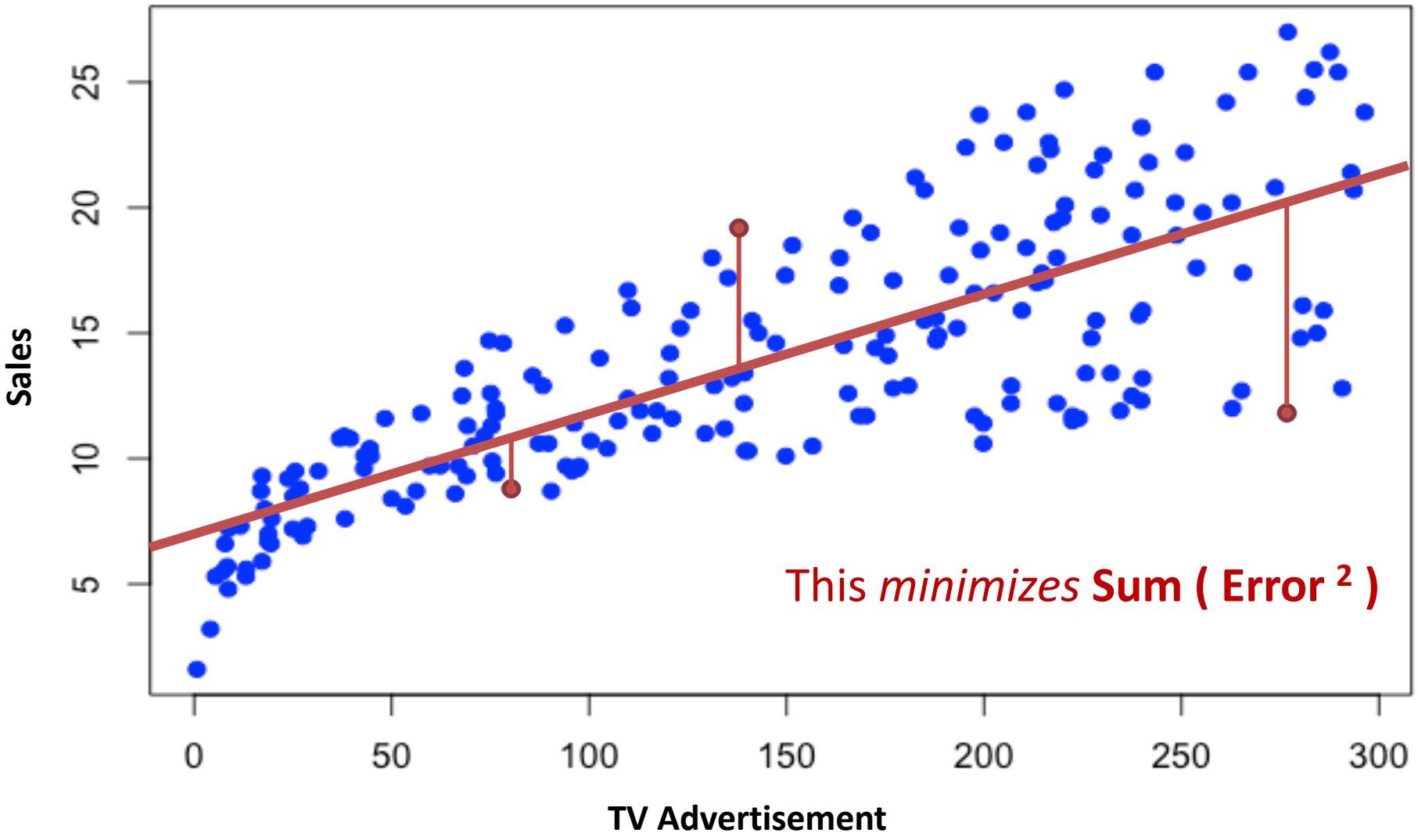
Distribution of Sales with respect to TV advertisement



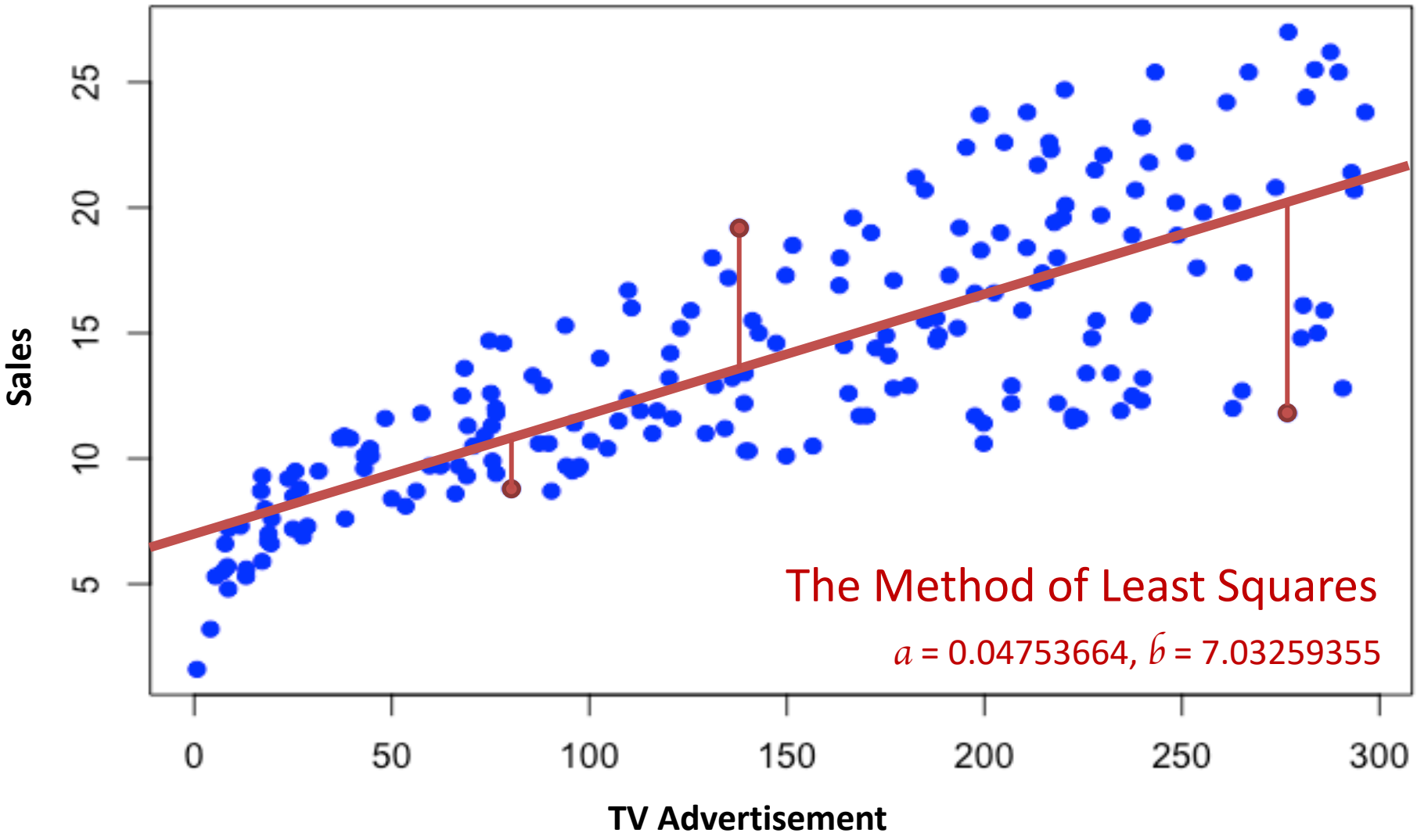
Distribution of Sales with respect to TV advertisement

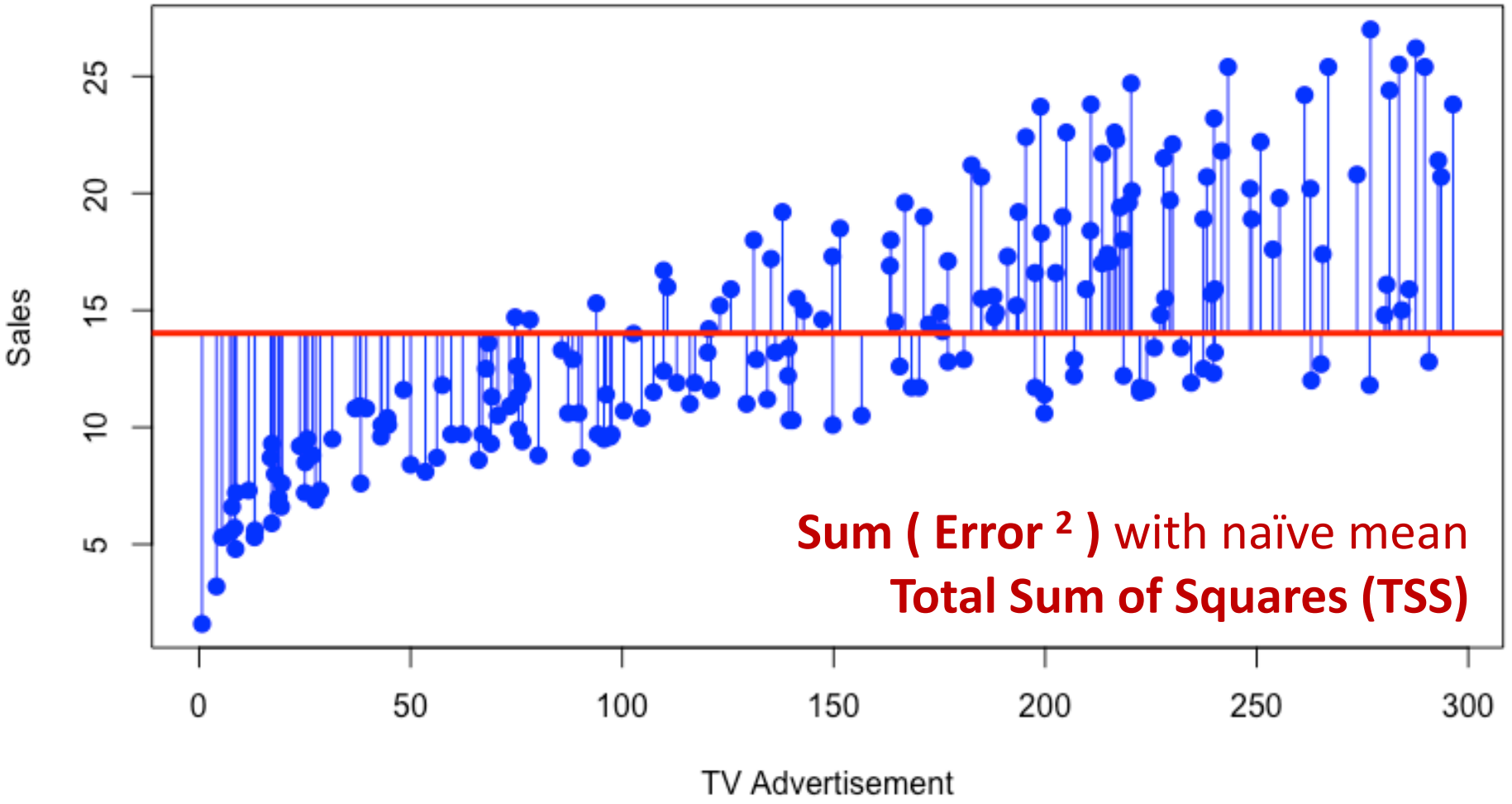


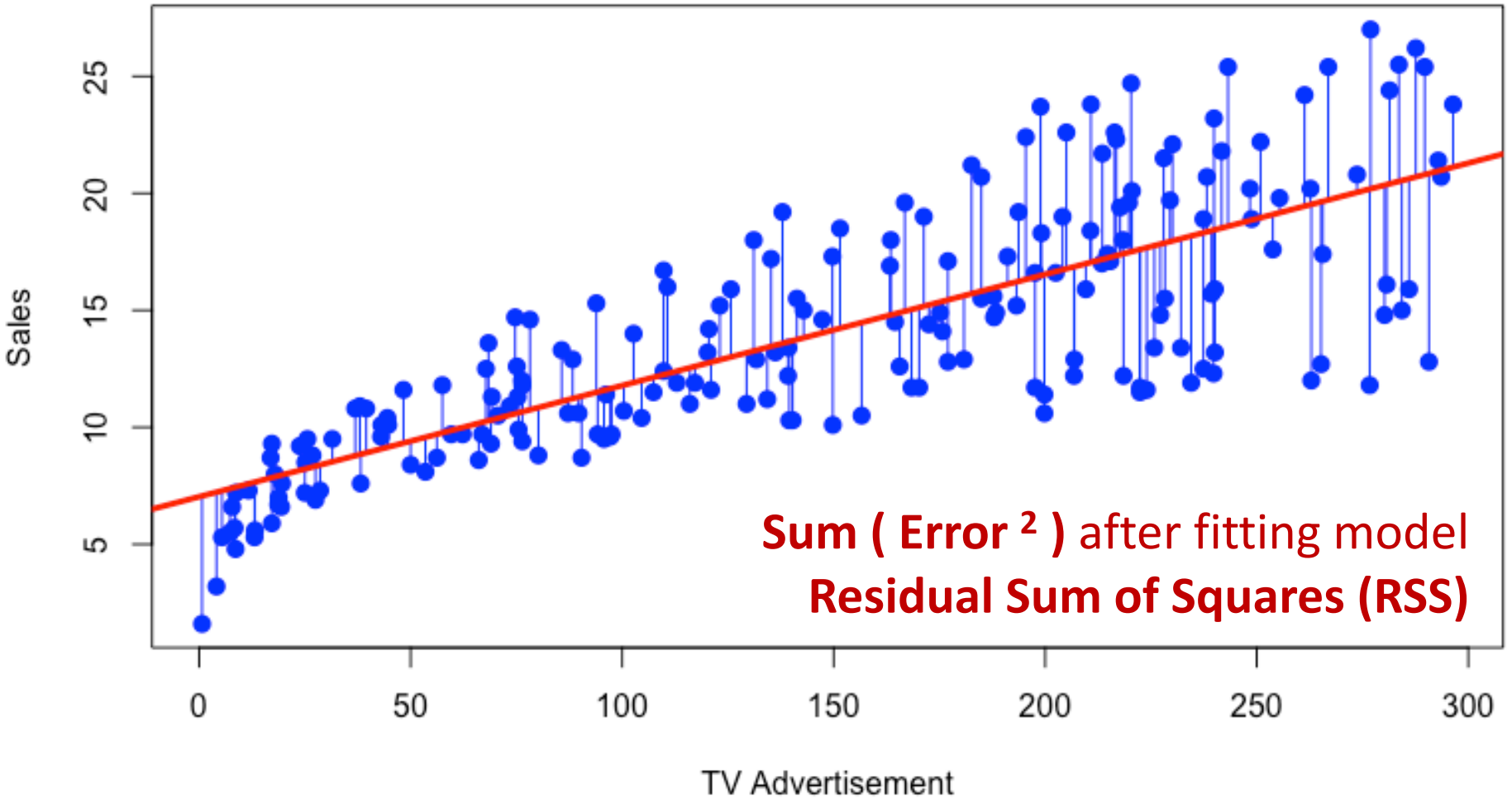
Distribution of Sales with respect to TV advertisement



Distribution of Sales with respect to TV advertisement




Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV Advertisement
Sum (Error ²) with naïve mean
Total Sum of Squares (TSS)


Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV Advertisement
Sum (Error 2) after fitting model
Residual Sum of Squares (RSS)

Sales Prediction

Given	:	Data from 200 *similar* stores of the Company
Data	:	Sales vs TV advertisement for *each* such Store
Strategy	:	Linear Regression (Sales vs TV advertisement)
Result	:	Obtained the *best-fit* (optimal) linear model by minimizing the 'Sum of Squares of Errors'
Best-fit line	:	**Sales =** a **x TV +** b with a = 0.0475, b = 7.0326
Prediction	:	**Sales = 0.0475 x TV + 7.0326**