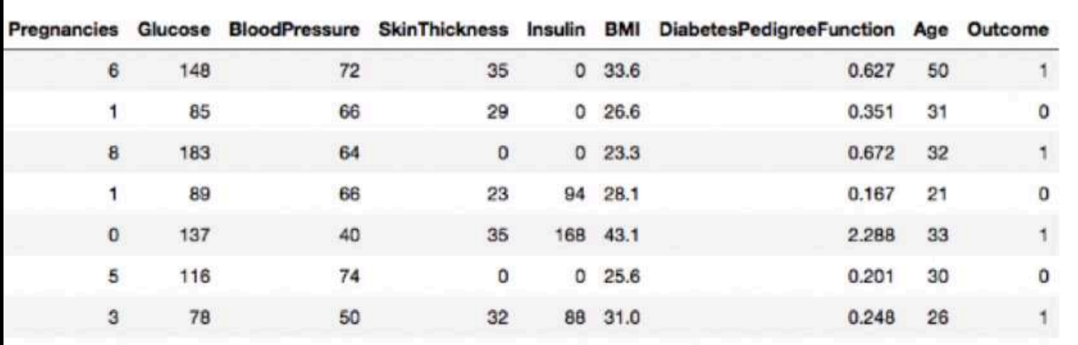
Agenda:

- Framework of problem statement in Data Science, Machine Learning, AI
- Different types of problems
- Different types of roles

About me:

- I love teaching
- Current:
  - Lead AI Scientist and Instructor at Scaler
- Past:
  - Lead AI Scientist at Target
  - AI Scientist at AlphaICs
  - PhD from Indian Institute of Science,
  - B Tech from NITK, Surathkal

Yes or no
↓



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31.0	0.248	26	1

Which attributes are most relevant

→ statistics / Correlation / hypothesis test

Can you predict whether a new person will have diabetes?

→ Machine Learning "SKLEARN" → Python

"Logistic regression"

"Decision tree"

"Random forest" "XGBoost"

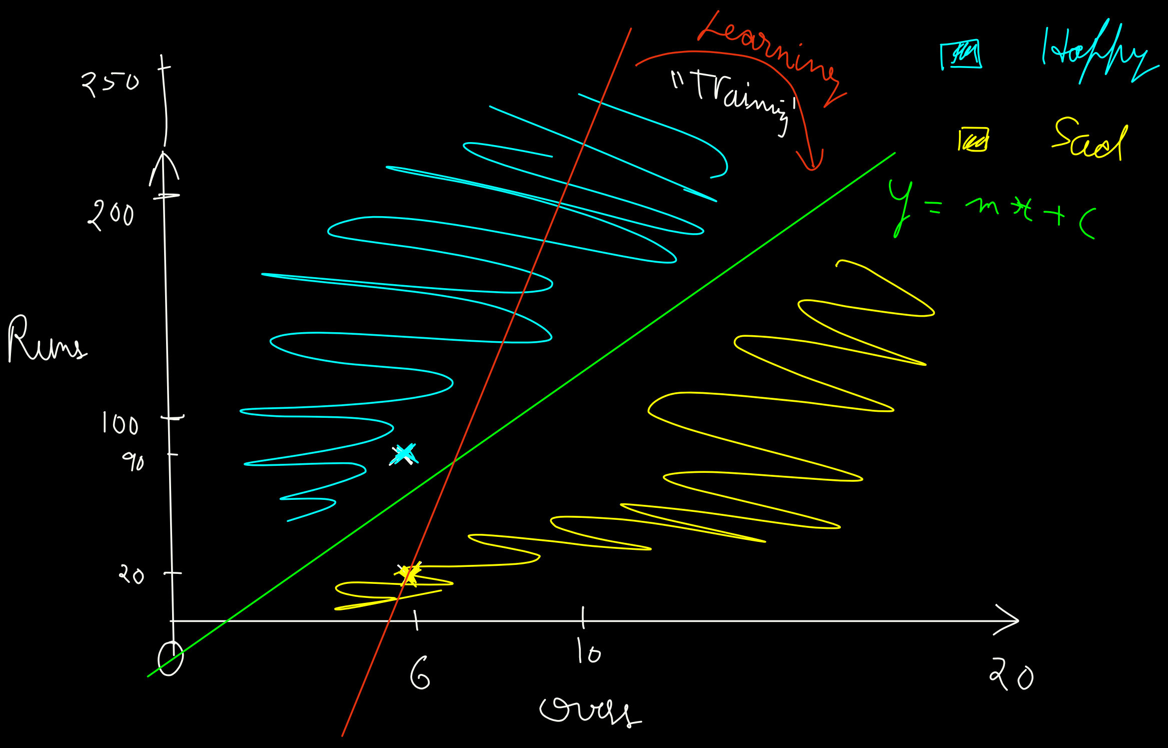
"Support vector Machine"

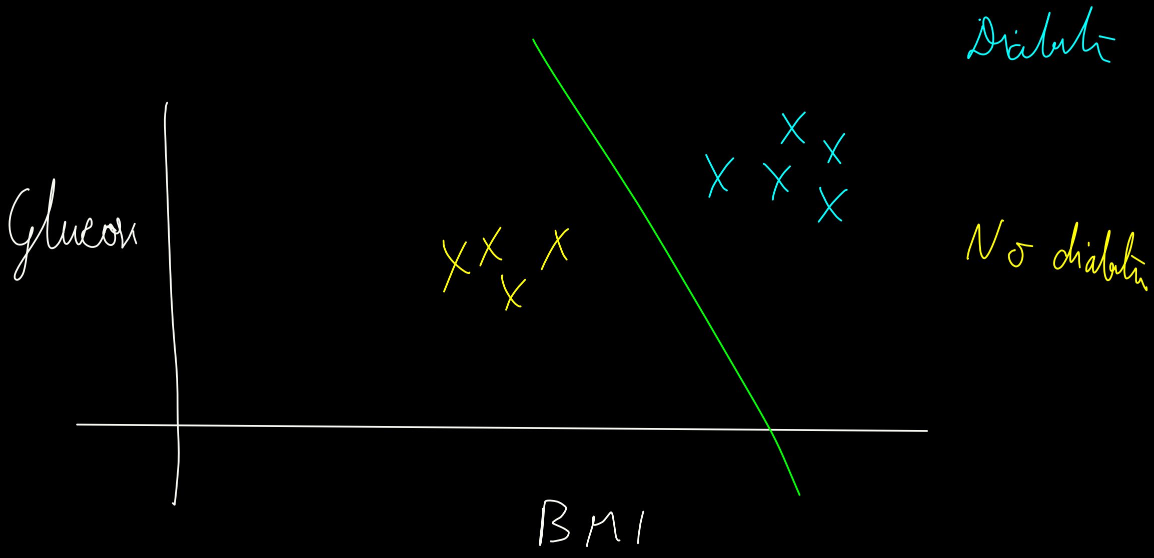
IPL Machine Learning

6 overs → 90 runs "Happy"

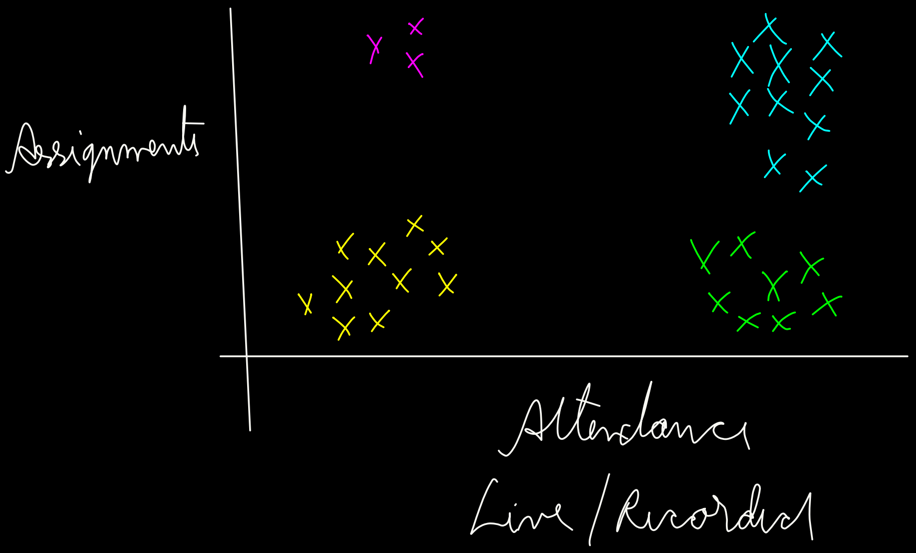
6 overs → 20 runs "Sad"

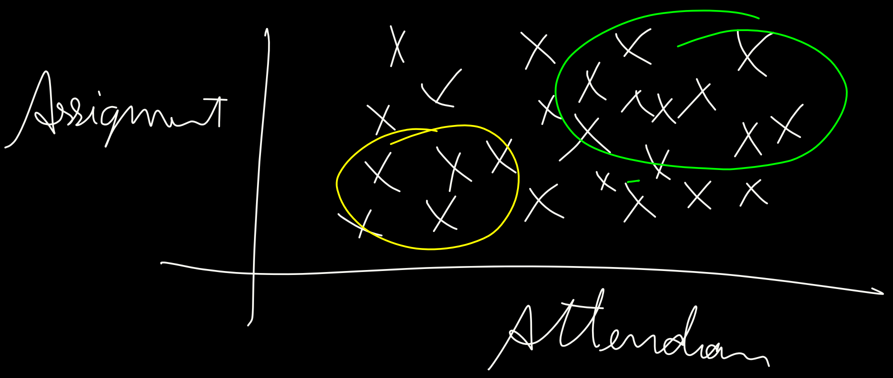
15 overs → 90 runs "Sad"





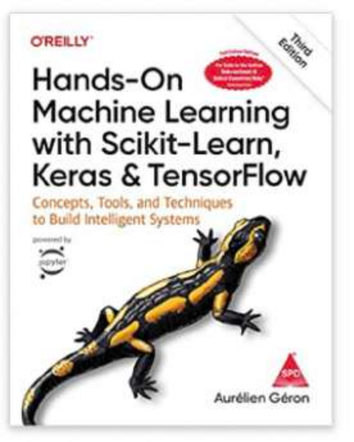
Lot of Students





"Clustering" K - means

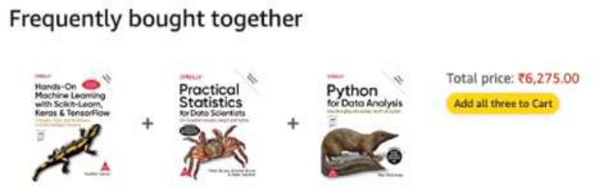
"Un supervised" cluster



Recy Sys

Recommends

↓

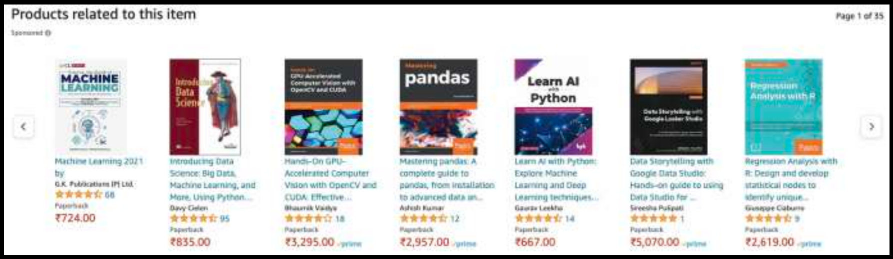

Frequently bought together

+ +

Total price: ₹6,275.00

Add all three to Cart

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
Practical Statistics for Data Scientists
Python for Data Analysis


Products related to this item

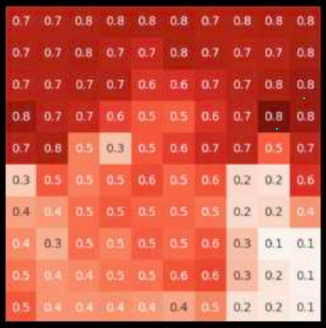
Sponsored

Page 1 of 35

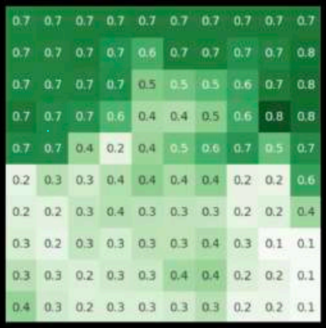
- Machine Learning 2021 by G.K. Publications (P) Ltd. ★★★★☆ 68 Paperback ₹724.00
- Introducing Data Science: Big Data, Machine Learning, and More, Using Python... Davy Cielen ★★★★☆ 95 Paperback ₹835.00
- Hands-On GPU-Accelerated Computer Vision with OpenCV and CUDA: Effective... Bhaumik Vaidya ★★★★☆ 18 Paperback ₹3,295.00 prime
- Mastering pandas: A complete guide to pandas, from installation to advanced data an... Ashish Kumar ★★★★☆ 12 Paperback ₹2,957.00 prime
- Learn AI with Python: Explore Machine Learning and Deep Learning techniques... Gaurav Leekha ★★★★☆ 14 Paperback ₹667.00
- Data Storytelling with Google Data Studio: Hands-on guide to using Data Studio for ... Sireesha Pulipati ★★★★★ 1 Paperback ₹5,070.00 prime
- Regression Analysis with R: Design and develop statistical nodes to identify unique... Giuseppe Ciaburro ★★★★☆ 9 Paperback ₹2,619.00 prime

increase sales

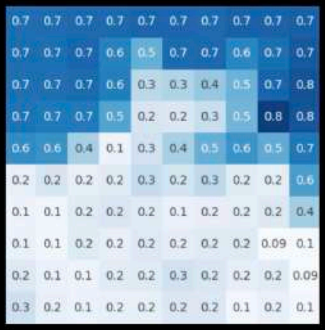
2012 → "AlexNet" ImageNet



0.7	0.7	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.8
0.7	0.7	0.8	0.7	0.7	0.8	0.7	0.7	0.7	0.8
0.7	0.7	0.7	0.7	0.6	0.6	0.7	0.7	0.8	0.8
0.8	0.7	0.7	0.6	0.5	0.5	0.6	0.7	0.8	0.8
0.7	0.8	0.5	0.3	0.5	0.6	0.7	0.7	0.5	0.7
0.3	0.5	0.5	0.5	0.6	0.5	0.6	0.2	0.2	0.6
0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.2	0.2	0.4
0.4	0.3	0.5	0.5	0.5	0.5	0.6	0.3	0.1	0.1
0.5	0.4	0.4	0.5	0.5	0.6	0.6	0.3	0.2	0.1
0.5	0.4	0.4	0.4	0.4	0.4	0.5	0.2	0.2	0.1

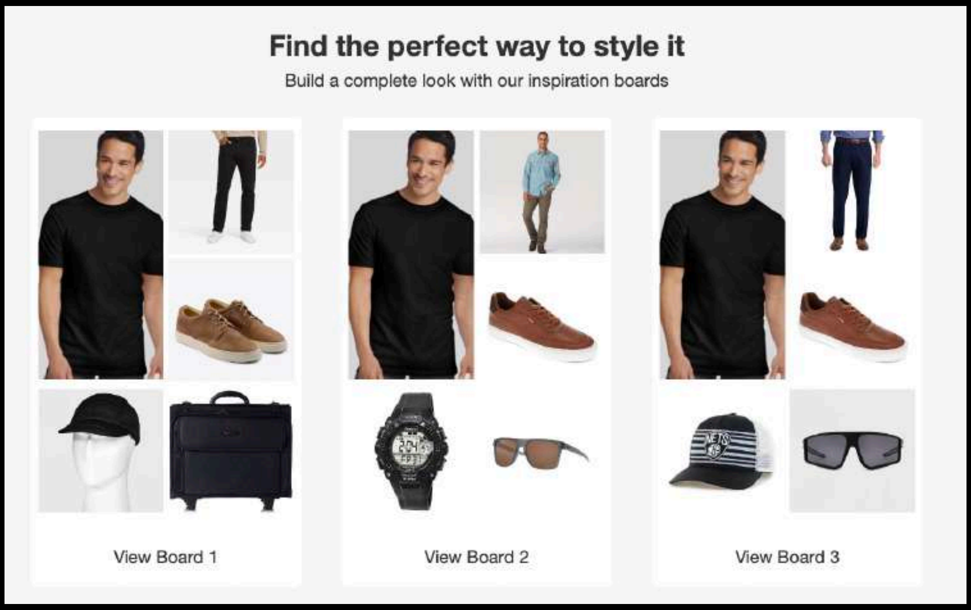


0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.7	0.8
0.7	0.7	0.7	0.7	0.5	0.5	0.5	0.6	0.7	0.8
0.7	0.7	0.7	0.6	0.4	0.4	0.5	0.6	0.8	0.8
0.7	0.7	0.4	0.2	0.4	0.5	0.6	0.7	0.5	0.7
0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.2	0.2	0.6
0.2	0.2	0.3	0.4	0.3	0.3	0.3	0.2	0.2	0.4
0.3	0.2	0.3	0.3	0.3	0.3	0.4	0.3	0.1	0.1
0.3	0.3	0.2	0.3	0.3	0.4	0.4	0.2	0.2	0.1
0.4	0.3	0.2	0.3	0.3	0.3	0.3	0.2	0.2	0.1



0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
0.7	0.7	0.7	0.6	0.5	0.7	0.7	0.6	0.7	0.7
0.7	0.7	0.7	0.6	0.3	0.3	0.4	0.5	0.7	0.8
0.7	0.7	0.7	0.5	0.2	0.2	0.3	0.5	0.8	0.8
0.6	0.6	0.4	0.1	0.3	0.4	0.5	0.6	0.5	0.7
0.2	0.2	0.2	0.2	0.3	0.2	0.3	0.2	0.2	0.6
0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.4
0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.09	0.1
0.2	0.1	0.1	0.2	0.2	0.3	0.2	0.2	0.2	0.09
0.3	0.2	0.1	0.2	0.2	0.2	0.2	0.1	0.2	0.1

"Deep Learning"



Find the perfect way to style it

Build a complete look with our inspiration boards

View Board 1

View Board 2

View Board 3

Segment Anything

Entry level

A 51%

- No python
- Not in SQL
- No Math

B 11%

- Yes python / Java / C++
- Not in SQL
- No math

C 16%

- Yes python
- Yes SQL
- No Math

22%

- Yes python
- Yes SQL
- Yes Math

A : Beginner → Target a Data Analyst role

B : Beginners w/o python

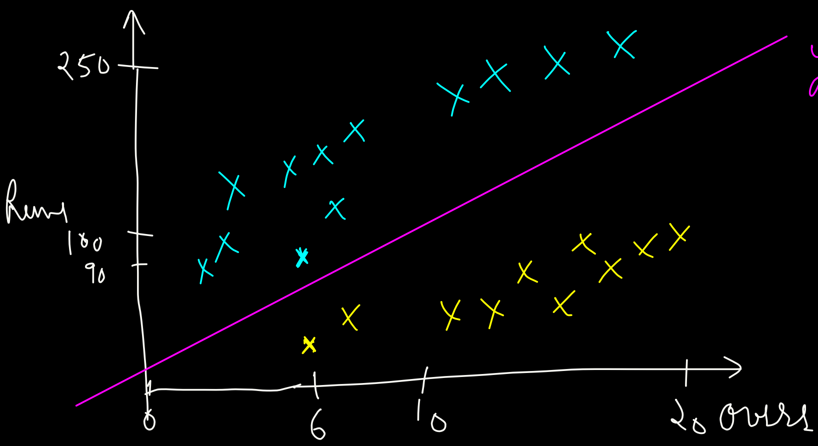
C : Intermediate

D : Advanced

IPL :

6overs → 90 runs Happy

6overs → 20 runs Sad



$y = \underline{m}x + \underline{c}$

① "Computer vision"

2012 → AlexNet

"Deep Learning"

VGG / ResNet

"Neural networks"

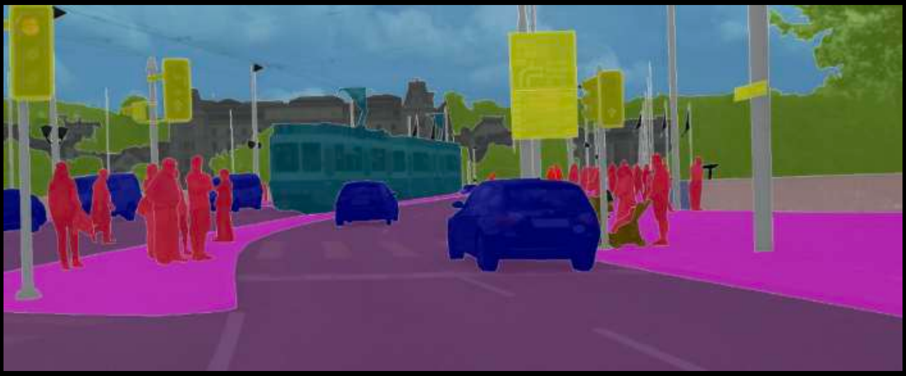
$$\begin{bmatrix} 25, 30, 40 \dots \\ \cdot \quad - \quad - - \end{bmatrix} \longrightarrow \text{"car"}$$

② "Natural Language Proc"

NLP

NLP

Audio → text NLP



Computer vision → Deep Learning

House prices   loc.   size, 1 or 2 or 3 BHK } → price

Machine Learning

Size	loc	Bed	Price
1000	Ba	3	1Cr

← tabular data

Eg: Linear regression
Decision Tree

Data science

↓

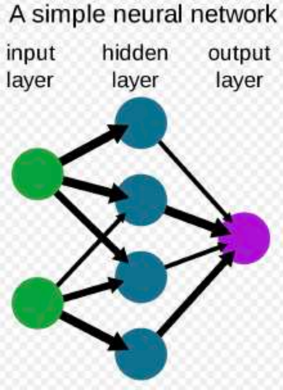
cricinfo data

	run	ball	4	6	SR	. . .	Win
1							✓
2							✗
3							✗
⋮							✓

Prob of Kohli scoring century?

Prob of winning when Kohli scores century

Deep learning



→ classify

Draw

Word-cloud

Null hypothesis

Test-stat

$\boxed{\text{P-value}} \rightarrow$ if $p < \alpha \rightarrow$ reject H_0 (Choose H_a)

"α" Significance level (0.05, 0.01)

Confusion matrix (FP, FN)

$P\left[\text{Data as extreme as that observed} \mid H_0 \text{ is true}\right]$

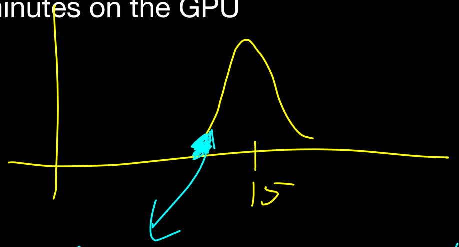
Burger company
Its burger weighs 200 grams
An unsatisfied customer, who is still hungry, wants to disprove this claim

Customer should prove that on average, the burger weighs < 200

H0: average = 200
Ha: average < 200

AI chip company
This company wants to claim that it is better than GPU
The training time for ResNet is 15 minutes on the GPU

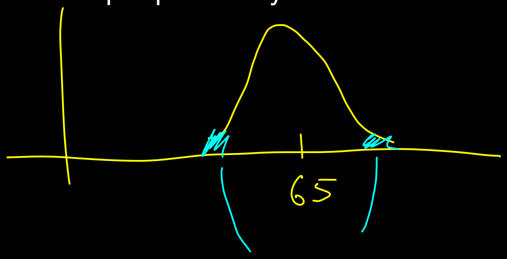
H0: training time = 15
Ha: training time < 15



left side (Ha: "<")
left tailed test

The average height of Indians is 65 inches
You want to verify whether this is true for people from your state

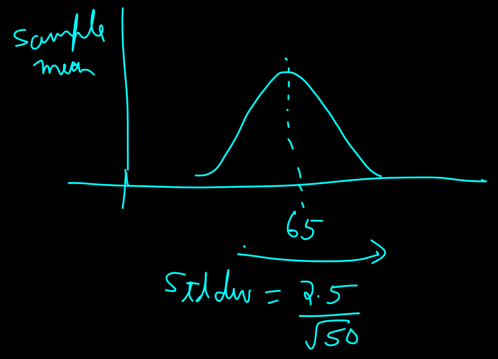
H0: height = 65
Ha: height != 65



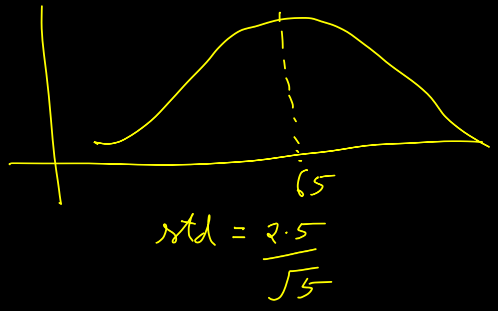
two-tailed test

Retail example

Recap of CLT
Average height is 65 inches, std dev 2.5
We sample 50 people
Let "m" denote sample mean
Is "m" a random variable? yes
What is its distribution? Gaussian (normal)
What is E[m]? 65
What is the std dev of m? $\frac{2.5}{\sqrt{50}}$



Sample 5 people
m → sample mean
$E[m] = 65$
Std dev = $\frac{2.5}{\sqrt{5}}$



Retail outlet with 2000 stores.
Weekly sales: Shampoo bottles: mean = 1800, std dev = 100

1)
We want to hire a marketing team to improve sales

Test them in 50 stores
In these 50 stores, our average sales is 1850
Want 99% confidence --> alpha = 0.01 (significance level)

2) Another team is deployed, and their average is 1900, number of stores was 5

What is the null hypothesis

H0: average = 1800 (marketing has no effect)
Ha: average > 1800 (marketing has effect)

Data: 50 stores, average here is 1850
Test statistic "m": sample mean of 50 stores
Distribution of test statistic? Gaussian
What is E[m] = 1800 (under H0)
What is std dev of m? 100/root(50)

$P(\text{extreme data} \mid H_0 \text{ is true})$

$P[m > 1850 \mid H_0 \text{ is true}]$

$1 - \text{norm.cdf}(3.53) = 0.0002 < 0.01$

Reject H_0 (Marketing had effect)

H_0: avg = 1800

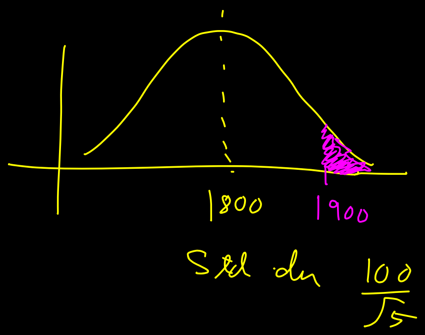
H_a: avg > 1800

5 - stores sample mean obs was 1900

Test statistic m = sample mean

$E[m] = 1800$

Std dev = $\frac{100}{\sqrt{5}}$



$P[m > 1900 \mid H_0 \text{ is true}]$

$$z = \frac{1900 - 1800}{100/\sqrt{5}} = 2.23$$

$1 - \text{norm.cdf}(2.23) = 0.012 > 0.01$

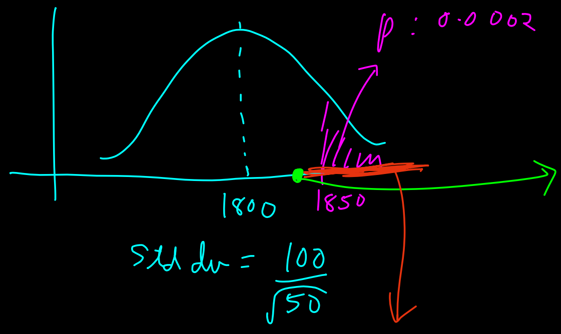
Effect is not statistically significant (stick H_0)

$\mu = 1800$
$\sigma = 100$

H_0 : avg $= 1800$
H_a : avg > 1800

50 story

$\alpha = 0.01$



Std dv $= \frac{100}{\sqrt{50}}$

z- score   norm.ppf(0.99) = 2.32

$$1800 + \frac{100}{\sqrt{50}}(2.32) = 1832.8$$

"critical value"

> 1832.8

"critical region"

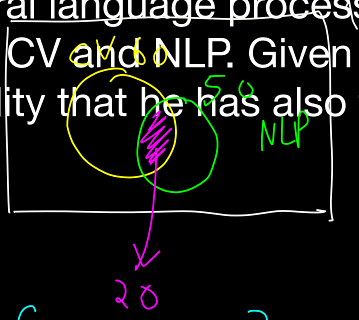
Interview

Style Class

Among 100 students, 60 have taken the computer vision (CV) module, 50 have taken natural language processing (NLP). Also, it is seen that 20 have taken both CV and NLP. Given that a person has taken NLP, what is the probability that he has also taken CV?

$P[CV] = \frac{60}{100}$

$P[NLP] = \frac{50}{100}$



$\frac{20}{100}$

① $P[CV \cap NLP]$

② $P[CV \mid NLP]$ ✓

③ $P[NLP \mid CV]$

$$P[CV \mid NLP] = \frac{P[CV \cap NLP]}{P(NLP)} = \frac{20/100}{50/100} = \frac{20}{50}$$