1. What is statistics
2. Types of statistics
3. Population vs sample
4. Data types
5. Levels of data
6. Parameter vs statistics
7. Frequency table
8. Frequency distribution table
9. Bar plot
10. Pie chart
11. Histogram
12. Distribution plot
13. Central tendency
14. Mean-median-mode
15. Mean vs median
16. Outlier
17. Types of distribution
18. Neg skew-pos skew-no skew
19. Range- mean deviation – absolute mean deviation
20. Variance
21. Standard deviation
22. How to find outlier
23. Percentile -Quartile
24. Box plot
25. Univariate-bivariate-multivariate
26. Scatter plot
27. Covariance
28. Covariance matrix
29. Correlation
30. Difference between covariance and correlation
31. Empirical rule for normal distribution
32. Chebyshev rule for not normal distribution

**Data standardization techniques:**

| **Age (years)** | **Income (rs)** | **Experience (years)** |
| --- | --- | --- |
| **30** | **50k** | **3** |
| **31** | **75k** | **6** |
| **32** | **1l** | **9** |
| **33** | **1.5l** | **12** |

One observation:

{Age:30,Income:50000,Experience:3}

3 columns are there

And 3 dimensions

Age is different data

Income Is different data

Experience is different data

All columns has different values and different units

Before apply any ML algorithm

It is very important all the data need to be under one layer

30 years ========== 50000========= 3

1. Z-standardization(z -score)
2. Normalization

They will change the values in such a way all the values under one roof

Age age\_z salary salary\_z

30 30-32/1=-2 50000 50000-100000/250000= -50000/25k=-2

31 31-32/1=-1 75000

32 0 100000

33 1 150000

34 2 200000

Mean = 32

Std = 1

1=

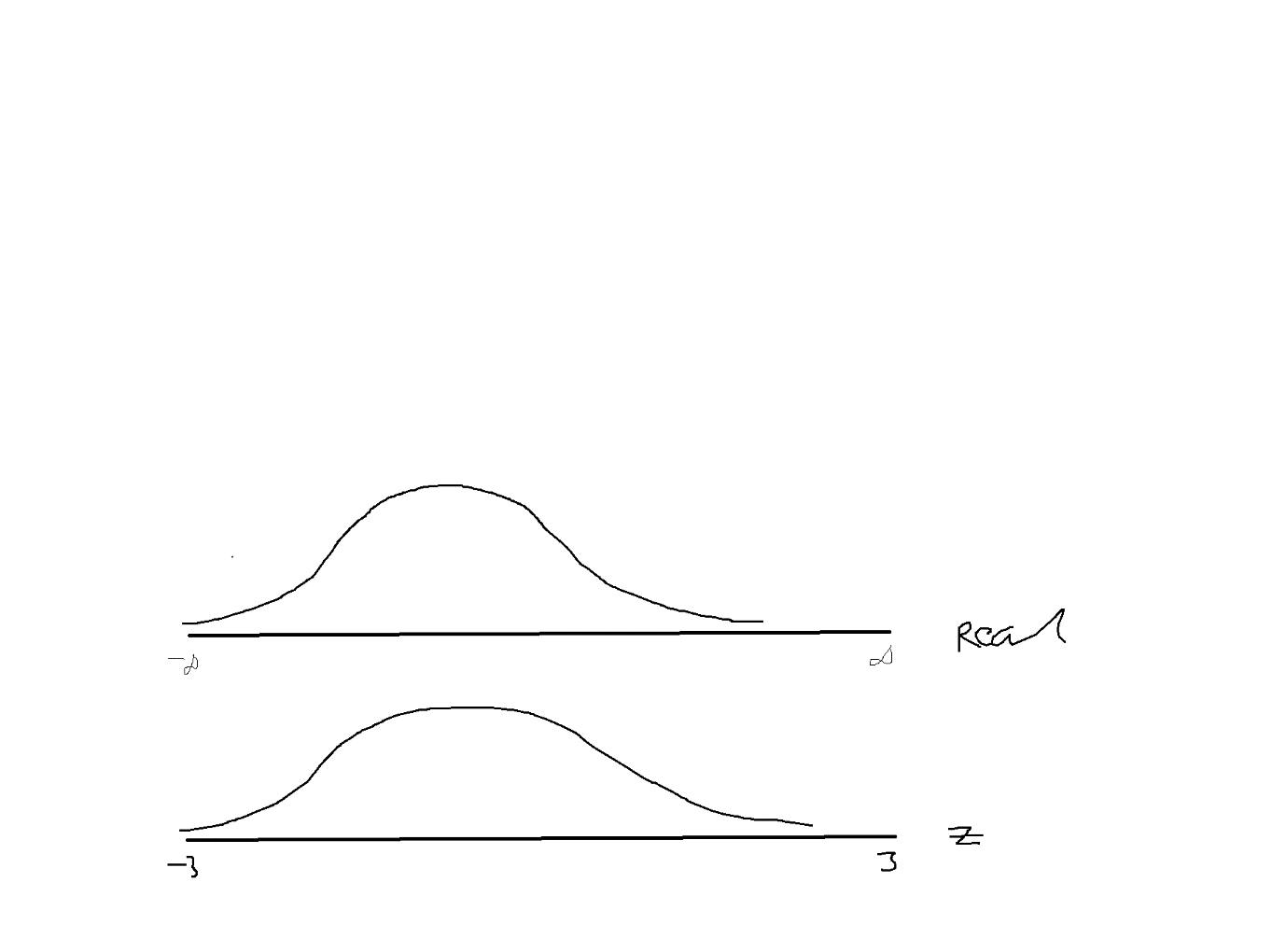
* Before every column has different units
* And different values
* Some column values might be very high
* Some column values are very less
* Our goal here is will convert all the values under one layer
* Z-standardization
* After you converting data, if you calculate mean of new data = 0
* After you converting data, if you calculate std = 1

Max value in original data below <3

Min value in original data above -3

The data range -3 to 3

Empirical rule says that 99.7% data lies -3 to 3 only



**Normalization:**

**Min max scaler**

**Suppose in one column**

**Max value**

**Min value**

1. **to 255**

**min= 0**

**max = 255**

**x-0/255= 255/255 =1**

**normalization means all the values below 1 only**

**in original data you found some outliers**

**if you want remove those outliers or not**

**is that outlier really an influential outlier or not?**

**Q3+3\*IQR Q1-3\*IQR**

**We need to change the scale of the data**

**Will standardize the data**

**Will change all the values of data in to a new line -3 to 3**

**Again will perform outlier analysis on the modified data**

**Still the same observation you found in new data ======== influential outlier**

**Age ============ age\_z (x-mu/sigma) -3 to 3**

**30 x1**

**31 x2**

**150 x3 : 2.3**

**150 is a outlier**

**It is really impact your data or not**

**You perform outlier analysis on new data**

**Q1 Q2 Q3**

**IQR**

**Q3+3\*IQR > outliers**

**If x3 again see as outlier**

**It is outlier in original data as well as standardized data**

**It is a influential outlier**

**How to treat outliers?**

1. **Remove the outliers**

**At what situation you will remove the outlier**

**100 observations**

**2 outliers ============== > 98**

**25 outliers =========== > 75**

**In order to train any ML model , we need more data**

**You are training your kid 5qns**

**Kid 20qns ======== > more scope to understand the patterns**

* **If your outliers are more , it is not good to remove completely**
* **Instead of removing replace outlier with median values**
* **Median does not affect by outlier**
* **You can replace by Q3 values ( capping)**
* **More than Q3 values relace with Q3 only**

**500 25000 50000 75000 1cr**

**Min Q1 Q2 Q3 max**

**Outliers = 75000 + 3\*50000 = 75000+ 1,50000= 2,25000**

**Those observation which are more than 2.25L is called as outlier**

**If that outliers more than 5% in your data**

**Replace those value with 50000 > 2.25lakhs ======= > 50k**

**Or**

**Cap those values with Q3 = 75000 >2.25lakhs ======= > 75k**

**If outliers : Q1-3\*IQR =========== > Q1 values**

**Cap with Q1 value**

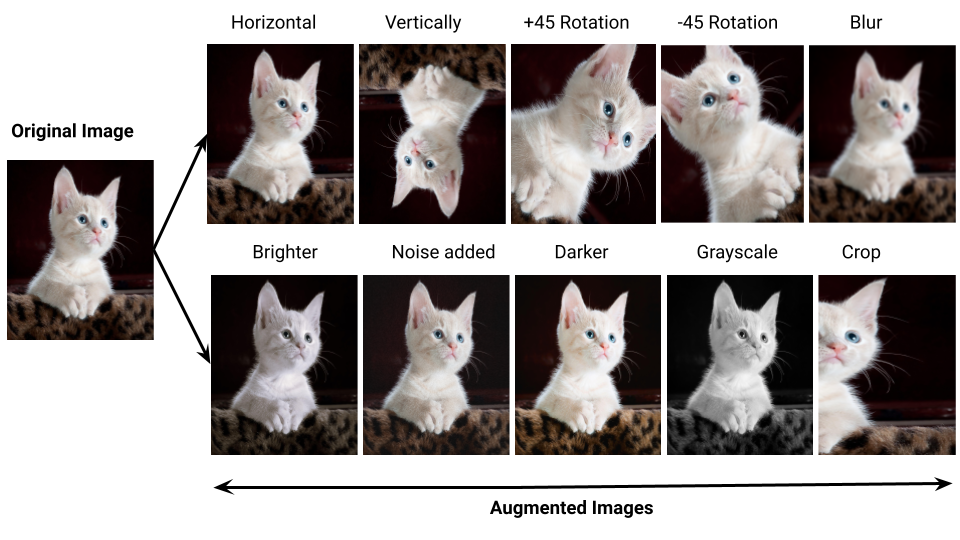
**He will take all outlier he will find median or mean , he can replace**

**Suppose your manager said**

**No I want to remove all the outliers, I don’t mind if you get less data**

**We need to create our own data**

***Data augmentation ========= DL NLP AI***



**Age gender income**

**30 M 50000**

**31 M 50000**

**30 F 50000**

**30 M 49999**

**50001**

**Influential outlier:**

In original data you might find some outliers

Ask yourself: Is that outliers really influential

Some outliers affect your data

Some outliers might not affect your data

In order to know that which are influential

We need to standardize the data, again we need find outlier for the new transformed

If both indexes

Suppose in original data 100th observation is outlier : 20000

In transformed data the same 100th observation coming as outlier

Then this 100th observation is called influential outlier

How to treat outliers:

If outliers percentage less than 2% ========== you can remove the outliers

Suppose data has total 100 observations , in that 2 observations are outliers

Remove those 2 observations

If it is more than 2% ============ do not remove

Simply replace those outliers with median value

Or

Replace with Q3 those values more than Q3

Replace with Q1 those values less than Q1

If you remove all outliers, then data will be reduce

Less data not enough to train ML model

Then use data augmentation tech

Z- data the mean always =0, std =1 the data lies between -3 to 3

**Convert Categorical data to Numerical data:**

* **We know that ML models developed by Maths**
* **Maths means allows only numbers**
* **So It is very very Important convert categorical data to numerical data**
* **Label encoder**
* **One hot encoder**

**Label encoder:**

**How many unique labels are there : 3**

**0 1 2**

**1 2 3**

| **Marital status** | **Marital status** |
| --- | --- |
| **Married** | **0** |
| **Unmarried** | **1** |
| **Divorced** | **2** |

**Married ================== 0**

**Unmarried ================ 1**

**Divorced ================ 2**

**In Visa data set**

**Case\_status**

**Denied =================== 0**

**Certified ================== 1**

**ML model output =========== 1 (certified )**

**=========== 0 (Denied)**

**One Hot Encoder:**

**How many unique labels : 3**

**3 separate columns**

**One hot : at a time only one will be ON , others will be OFF**

| **Marital status** | **Marital\_status\_married** | **Marital\_status\_unmarried** | **Marital\_status\_divorced** |
| --- | --- | --- | --- |
| **Married** | **1** | **0** | **0** |
| **Unmarried** | **0** | **1** | **0** |
| **Divorced** | **0** | **0** | **1** |

**421 code**

| **4** | **2** | **1** | **Nu** |
| --- | --- | --- | --- |
| **0** | **0** | **0** |  |
| **0** | **0** | **1** |  |
| **0** | **1** | **0** |  |
| **0** | **1** | **1** |  |
| **1** | **0** | **0** |  |
| **1** | **0** | **1** |  |
| **1** | **1** | **0** |  |
| **1** | **1** | **1** |  |

In One hot encoder the biggest advantage is

The new columns are independent each other

There is a scenario if two columns are dependent ====== > ML model will not able to understand which column is important

In one hot encoder at a time only one is ON, others will be OFF

* **Independent each other**
* **Perpendicular each other**
* **Orthogonal each other**
* **90 degrees phase shift**

**Disadvantage:**

**Suppose if you have 100 unique labels are there**

**100 new columns will be created**

**Data dimensions will be increased**

**When you pass the ======== more time/complexity**

**Curse of Dimensionality : we are increasing the dimensions of the data**

**196 counties are codes ========== > 1 to 196**

**Manuly**

**IN: 1**

**US:2**

**How to convert Normal distribution**

**How to select important features**

**PCA**

**Hi sir,  
What techniques will be followed for exit polls results. usually they will collect sample data right.  
Is there any data science concepts will apply here like data augmentation.**

**Prepoll**

**Postpoll**

**Data drift =======**

**Before 10 yaers ============ data**