Tutorial: 2 U18C0018

DWDM

1→ Needs of Pre-Processing of data

⇒ As data in the real world is dirty

⇒ Incomplete : lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

eg occupation = " "

– Noisy : containing error/ outliers

eg Salary : "-10"

– inconsistent: containing discrepancies in codes/names

eg Age = "42" Birthday = "03/07/1997"

eg was rating "1,2,3", now rating "A,B,C"

eg discrepancy b/w duplicate records

2→ A well accepted multidimensional view of data quality

→ Accuracy → Interpretability

→ Completeness → Acessibility

→ Consistency

→ Timeliness

→ Believability

3→ Major tasks in data Preprocessing:

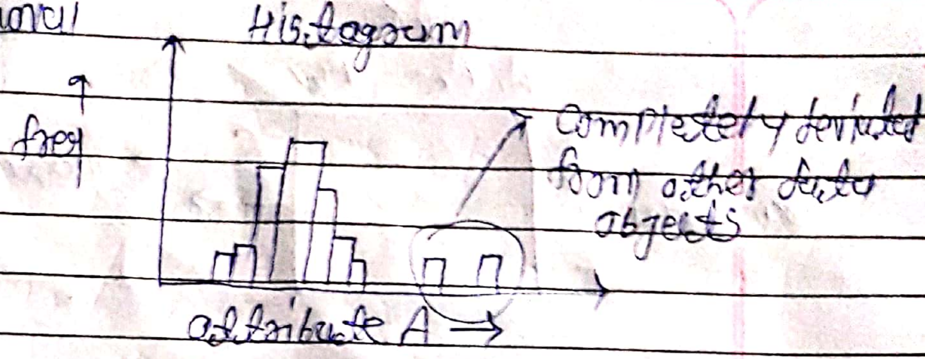
(1) Data Cleaning: also known as scrubbing

This task involves filling of missing values smoothing/ removing noisy data & outliers along with resolving inconsistencies

EX missing values

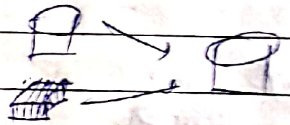
age	income	student	buys-computer
<=30	high	no	?
>40	medium	yes	?

Outliers Removal


Histogram

(2) Data integration:- This task involves integrating data from multiple sources such as databases, data cubes, files, etc. The data sources can be homogenous / heterogeneous. The data obtained from the sources can be structured, unstructured / semi-structured format.

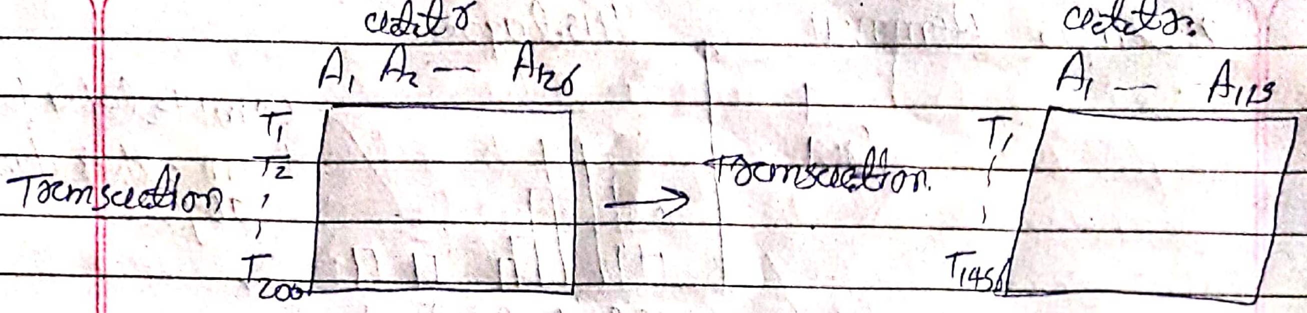
Ex Attribute of customer identification may be referred to as customer id in one data store & cus-id in other. Naming consistencies occurs. so this task is important.



(3) Data transformation:- Involves normalisation & aggregation of data according to needs of data set. (additional procedure)

Ex 2,32,100,59,48 → -0,02, 0.32, 1.00, 0.59, 0.48

(4) Data reduction:- Data is reduced. No. of records / no. of attributes / dimensions can be reduced. Done by keeping in mind that reduced data should produce same results as original data.



(5) Data discretization: considered as a Part of data reduction. The numerical attributes replaced with normal ones.

Ex age (numerical attribute) replaced by interval labels (0-10, 11-20 & so on) or conceptual labels (eg youth, adult & senior)

4→ Data Cleaning method of filling in missing values

we handle this by
(1) Ignore tuple
(2) fill in manually
(3) use a global constant to fill
(4) use attr values mean
(5) use most Probable value to fill
(6) use regression methods

Solution: ⇒ Deletion
⇒ Imputation

In Imputation
⇒ single value imputation.
replace missing value with a single value, mean, median, mode, mean Person of corres-Ponding feature.

→ Dataset that have great outliers, median Preferred.

→ for encoded categorical features use mode

→ Hot/cold deck imputation (randomly choose one of their values to fill in)

→ Imputation using KNN (feature similarity)

→ regression based imputation (simple, linear, stochastic, log-linear)

→ multiple imputation (several imputed values to each missing value)

5→ Measures of Center tendency

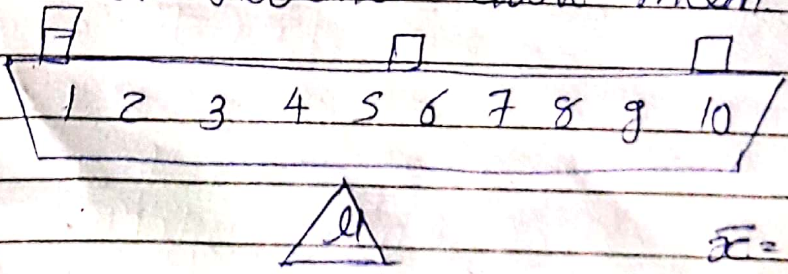
⇒ It is a statistical measure that determines a single values that accurately describes the center of distribution & represents the entire distribution of scores.

→ goal is to identify single value that is best representative for entire set of data

(a) mean : $\mu = \frac{\Sigma x}{N}$

→ It is an algebraic measure

→ Balance Point of distribution because the sum of distances below the mean is exactly equal to sum of distances above mean.

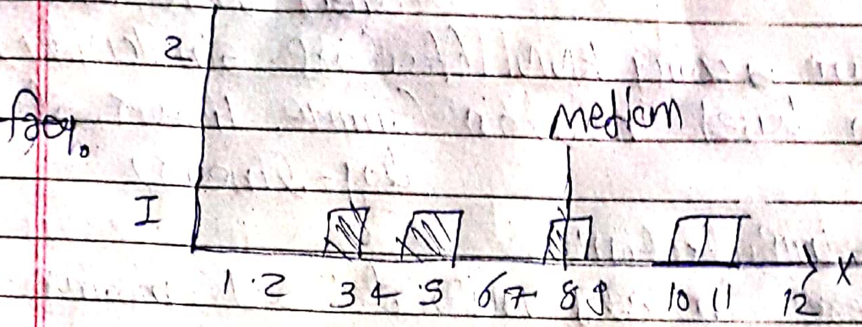


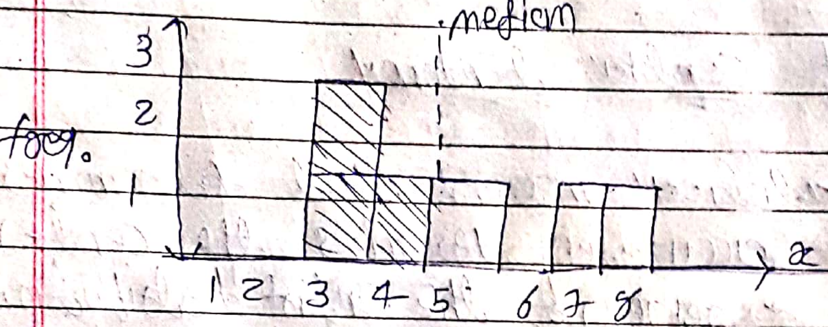
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

weighted mean

(b) median: If scores in a distribution are listed in order from smallest to largest, the median is defined as midpoint of list.

Ex





(c) Mode: most freq. occurring category/score in the distribution.

Ex 1, 2, 2, 2, 3, 3, 4, 5, 5

mode 2

→ These 3 are often systematically related to each other.

→ In symmetric distribution, mean & median will always be equal.

In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os
import warnings
```

In [2]:

```
warnings.filterwarnings("ignore")
```

In [3]:

```
df = pd.read_csv('C:\\Users\\91810\\Downloads\\Histograms.csv')
df.head()
```

Out[3]:

	A	B	C	D	Left Skew	Multimodal	IQ20	IQ100
0	48.916926	67.223785	55.917225	45.561471	23.1	37.632318	120.459951	93.041368
1	47.692726	68.175751	30.174288	47.825783	18.2	49.244001	107.418864	93.806158
2	48.629579	61.753451	43.641583	59.699370	14.6	37.780203	95.006312	135.339681
3	58.544034	69.783507	53.738745	45.704638	21.2	56.827208	96.522192	100.772632
4	44.821338	70.730153	67.829659	44.254419	24.5	54.513731	108.878563	91.600053

In [4]:

```
df.describe()
```

Out[4]:

	A	B	C	D	Left Skew	Multimodal	IQ20	IQ100
count	100.000000	100.000000	100.000000	100.000000	92.000000	200.000000	20.000000	100.000000
mean	50.632133	65.544513	50.851334	50.211539	20.107609	59.734576	102.132401	102.925179
std	5.063123	5.085469	15.342335	5.228720	7.047410	11.513170	15.550922	15.223586
min	39.935450	54.142510	15.381702	39.081231	1.000000	33.555815	78.284920	69.763146
25%	47.693309	61.819282	42.188371	46.852570	15.025000	49.592572	91.681628	92.096983
50%	50.673711	65.898797	51.654882	49.726685	21.500000	60.602041	105.608402	101.426575
75%	53.820237	68.821663	61.308291	53.196049	25.925000	69.521137	108.952938	114.041076
max	63.531483	80.184730	90.095257	71.200000	31.400000	81.929535	133.448312	138.871933

In [5]:

```
'''
Data Cleaning
Filling in missing values
Single value imputation -mean
'''
mean_val = df['A'].mean()
df_mean = df
df_mean['A'].fillna(value=mean_val, inplace=True)
df_mean.isna().sum()
```

Out[5]:

```
A               0
B             100
C             100
D             100
Left Skew     108
Multimodal      0
IQ20          180
IQ100         100
```