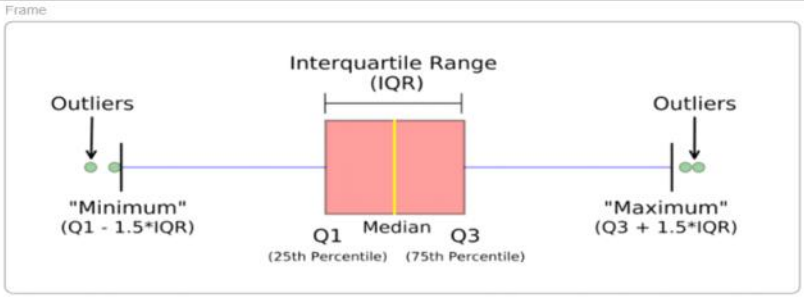
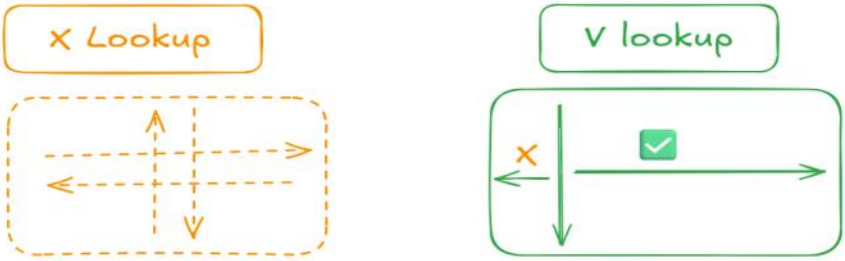
Data Cleaning in Excel (Handling Missing Values & outliers)

SESSION OVERVIEW

- How and when to use VLOOKUP, HLOOKUP, XLOOKUP, and INDEX-MATCH
- Check and understand data types in Excel
- Identify missing values
- Apply different techniques to handle missing data effectively





X-Lookup

X-Lookup,V-Lookup

TABLE 3		
CITY	EMPLOYEE ID	STATE
Columbus	110608	Ohio
Chicago	253072	Illinois
Tampa Bay	352711	Florida
Chicago	391006	Illinois
Chicago	392128	Illinois
Tampa Bay	549427	Florida
Columbus	580622	Ohio
Tampa Bay	602693	Florida
Austin	611810	Texas
Tampa Bay	612235	Florida
Austin	795574	Texas
Chicago	830385	Illinois
Austin	990678	Texas

"provide the column where you need to return the answer"

=XLOOKUP(110608 Doe

XLOOKUP(**lookup_value**, lookup_array, return_array, [if_not_found], [match_mode], [search_mode])

"NA/Not Found"

only a list or a column referencing the lookup column

```
=XLOOKUP(A3,'VLOOKUP(Different sheet)-2'!$G$3:$G$15,'VLOOKUP(Different sheet)-2'!$F$3:$F$15,"NA")
```

G3 =XLOOKUP(A3,'VLOOKUP(Different sheet)-2'!$G$3:$G$15,'VLOOKUP(Different sheet)-2'!$H$3:$H$15)

	A	B	C	D	E	F	G	H	I
1	TABLE 1								
2	EMPLOYEE ID	LAST_NAME	FIRST_NAME	CITY	STATE	CITY_X	STATE_X		
3	110608	Doe	John	Columbus	Ohio	Columbus	Ohio		
4	253072	Cline	Andy	Chicago	Illinois	Chicago	Illinois		
5	352711	Smith	John	Tampa Bay	Florida	Tampa Bay	Florida		
6	391006	Pan	Peter	Chicago	Illinois	Chicago	Illinois		
7	392128	Fawre	Bret	Chicago	Illinois	Chicago	Illinois		
8	549427	Elway	John	Tampa Bay	Florida	Tampa Bay	Florida		
9	580622	Manning	Eli	Columbus	Ohio	Columbus	Ohio		
10	602693	Vick	Micheal	Tampa Bay	Florida	Tampa Bay	Florida		
11	611810	Woods	Tiger	Austin	Texas	Austin	Texas		
12	612235	Jordan	Micheal	Tampa Bay	Florida	Tampa Bay	Florida		
13	795574	Stark	Tony	Austin	Texas	Austin	Texas		
14	830385	Williams	Price	Chicago	Illinois	Chicago	Illinois		
15	990678	Pitt	Brad	Austin	Texas	Austin	Texas		

H-Lookup

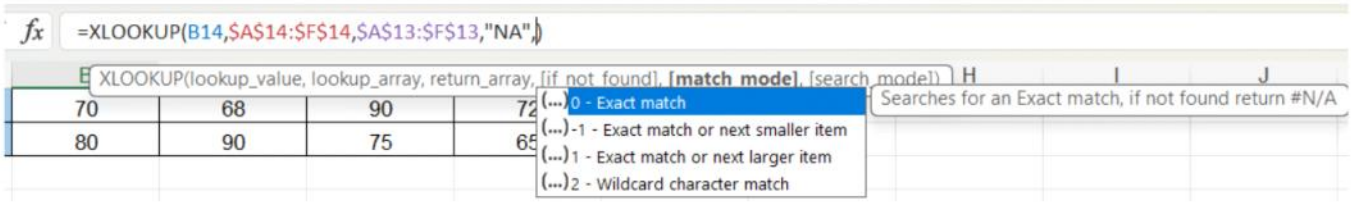
ACCOUNTS	75	65	70	60	59
ECONOMICS	65	72	78	89	67
STUDENT NAME	**A**	**B**	**C**	**D**	**E**
MANAGEMENT	70	68	90	72	58
MATHEMATICS	80	90	75	65	87

HARMEAN =XLOOKUP(B8,$B$15:$F$15,$B$14:$F$14,"Not Found")

XLOOKUP(lookup_value, lookup_array, **return_array**, [if_not_found], [match_mode], [search_mode])

	A	D	E	A	C	B
8		D	E	A	C	B
9	MATHEMATICS	72	58	70	90	68
10	ECONOMICS_X	$F$14,"Not	67	65	78	72
11						
12						
13	ACCOUNTS	75	65	70	60	59
14	ECONOMICS	65	72	78	89	67
15	**STUDENT NAME**	**A**	**B**	**C**	**D**	**E**
16	MANAGEMENT	70	68	90	72	58
17	MATHEMATICS	80	90	75	65	87

Match Mode

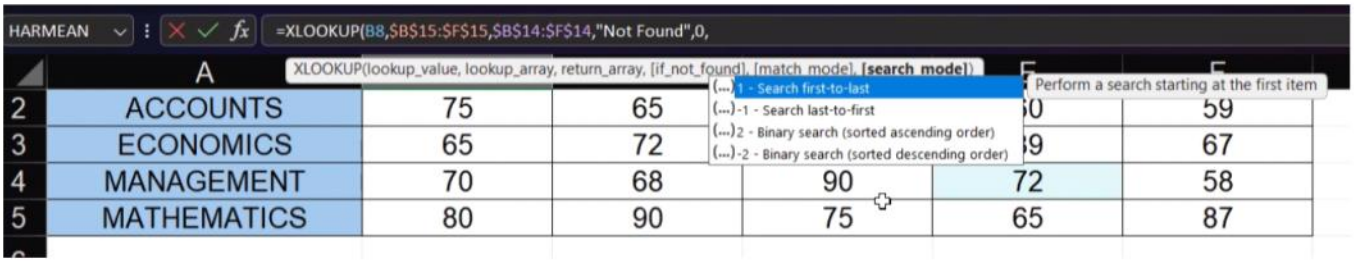


0-19 : Poor Performance
21-39 : Below Average
41-59 : Average
61-79 : Good
81-89 : Excellent
91-99 : Outstanding

Student Marks

20
40
35
55
90
100

Search Mode



First-to-last [searching top to bottom one by one]

Last-to-First [searching bottom to top one by one]

Linear Search

1 : Search First to Last
-1 : Search Last To First

24,1 [Default]

24,-1

10,12,15,9,10,15,12,4,5,67,12,34,24,74,23,24,6,343,56,42,25,64

Un-Sorted

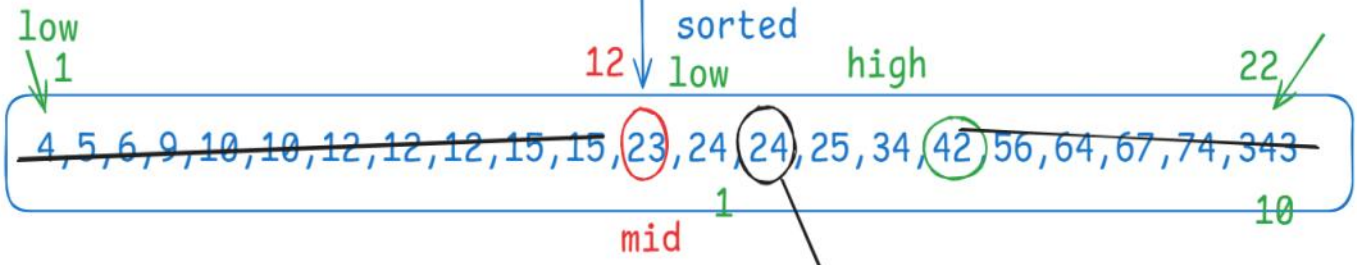
Time Complexity: O(N)

Binary Search

Sorted

Faster than Linerar Search

10,12,15,9,10,15,12,4,5,67,12,34,24,74,23,24,6,343,56,42,25,64



2 : Binary Search in Ascending
-2 : Binary Search in Descending

search in binary Search

[low] arr[0] = 4
[high] arr[length-1] = 343

Search Key:24

low = 1
high = 22
mid = (low + high)/2
= (1+22)/2 = 11.5 -> 12

Time Complexity: O(logN)

O(logN) > O(N)

INDEX - MATCH

	Height	Weight
Amanda	=INDEX(	

INDEX(**array**, row_num, [column_num])
INDEX(**reference**, row_num, [column_num], [area_num])

NAME	HEIGHT	WEIGHT
Sally	6.2	95
Tom	5.9	87
Kevin	5.8	88
Amanda	5.5	79
Carl	6.1	101
Ned	6	83

	Height	Weight
Amanda	=INDEX($A$1:$C$7,5,2)	=INDEX($A$1:$C$7,5,3)

Amanda [5],
Height [2]

Amanda[5]
Weight[3]

=MATCH(

MATCH(**lookup_value**, lookup_array, [match_type])

	Height	Weight
Amanda	5.5	79

Amanda	5
Height	2
Weight	3

=MATCH(A13,$A$1:$A$7,0)
=MATCH(B12,$A$1:$C$1,0)
=MATCH(C12,$A$1:$C$1,0)

Data Cleaning in Excel

- Missing Value
- Formatting Issue

Name	gender	math score	reading scor	writing score
Nancy	female	59	70	78
Anshul	male	96	93	87
Rani	female	57	76	77
Aryan	male	70	70	63
Anjali	female	83	85	86
Ram	male	68	57	54
Megha	female	82	83	80
Rosy	female	46	61	58
Shyam	male	80	75	73
Geeta	female	57	69	77
Varun	male	74	69	69
John	male	53	50	49
Harshit	male	76	74	76
Tarun	male	70	73	70
Taran	male	55	54	52
Tim	male	56	46	43

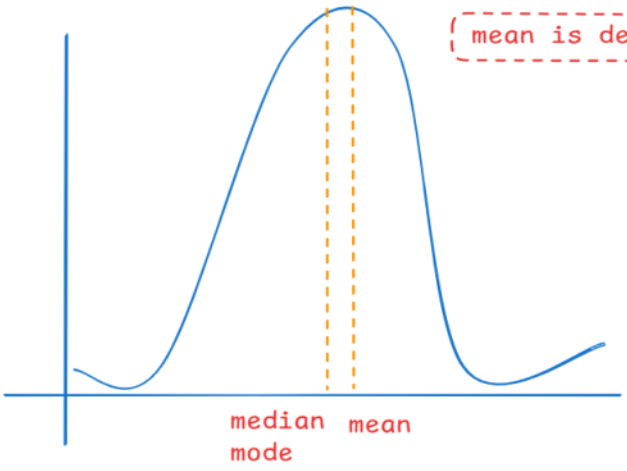
FORMULA	RESULTS	
AVERAGE=AVERAGE(Range of cells)	69.72727273	MATHS SCORE
	72.05050505	READING SCORE
	70.63636364	WRITING SCORE
MEDIAN=MEDIAN(Select the range of cells)	70	MATHS SCORE
	74	READING SCORE
	74	WRITING SCORE
MODE=MODE(Select the range of cells)	80	MATHS SCORE
	82	READING SCORE
	78	WRITING SCORE
STANDARD DEVIATION=STDEV(Select the range of cells)	15.2139961	MATHS SCORE
	14.42459191	READING SCORE
	14.74438983	WRITING SCORE
VARIANCE= VAR(Select the range of cells)	231.4656772	MATHS SCORE
	208.0688518	READING SCORE
	217.3970315	WRITING SCORE

FORMULA	RESULTS	
AVERAGE=AVERAGE(Range of cells)	=AVERAGE(C2:C100)	MATHS SCORE
	=AVERAGE(D2:D100)	READING SCORE
	=AVERAGE(E2:E100)	WRITING SCORE
MEDIAN=MEDIAN(Select the range of cells)	=MEDIAN(C2:C100)	MATHS SCORE
	=MEDIAN(D2:D100)	READING SCORE
	=MEDIAN(E2:E100)	WRITING SCORE
MODE=MODE(Select the range of cells)	=MODE.SNGL(C2:C100)	MATHS SCORE
	=MODE.SNGL(D2:D100)	READING SCORE
	=MODE.SNGL(E2:E100)	WRITING SCORE
STANDARD DEVIATION=STDEV(Select the range of cells)	=STDEV.S(C2:C100)	MATHS SCORE
	=STDEV.S(D2:D100)	READING SCORE
	=STDEV.S(E2:E100)	WRITING SCORE
VARIANCE= VAR(Select the range of cells)	=VAR.S(C2:C100)	MATHS SCORE
	=VAR.S(D2:D100)	READING SCORE
	=VAR.S(E2:E100)	WRITING SCORE

day	maxtemp	temparature
1	19.9	18.3
2	21.7	18.9
3	20.3	19.3
4	22.3	20.6
5	21.3	20.7
6	24.3	20.9
7	21.4	18.8
8	21	18.4
9	18.9	18.1

Mean	16.89393939
Median	16.6
Mode	16.6
Variance	12.98139147
Standard deviation	3.602969812

C	D	E	F
temparature			
18.3			
18.9			
19.3			
20.6		Mean	=AVERAGE(C2:C100)
20.7		Median	=MEDIAN(C2:C100)
20.9		Mode	=MODE.SNGL(C2:C100)
18.8		Variance	=VAR.S(C2:C100)
18.4		Standard deviation	=STDEV.S(C2:C100)
18.1			



mean is deviated towards right -> right Skewed[+ve Skewed]

temparature
18.3
18.9
19.3
200
300
20.9
18.8
18.4

Mean	21.52727273
Median	16.6
Mode	16.6
Variance	1150.974657
Standard deviation	33.9260174

Coefficient of Variation [CV] = SD/Mean * [100]

IF CV > 50% -> USE Median
IF CV <= 50% -> USE 'MEAN'

Mean	16.87835052
Median	16.5
Mode	16.6
Variance	13.23713058
Standard deviation	3.638286765
CV	22%
Which one to choose?	Mean

Mean	=AVERAGE(C2:C100)
Median	=MEDIAN(C2:C100)
Mode	=MODE.SNGL(C2:C100)
Variance	=VAR.S(C2:C100)
Standard deviation	=STDEV.S(C2:C100)
CV	=F9/F5
Which one to choose?	=IF(F10>50,"Median","Mean")

Moving Average [3 value]

SQL

NTILE() --> [5 moving Average]
[Ranges BETWEEN 2 preceding
AND 2 FOLLOWING]

Months	Temperature	
Jan	39	
Feb	42	43.7
Mar	50	50.7
Apr	60	59.8
May	69.5	69.5
Jun	79	77.8
Jul	85	81.7
Aug	81	82.5
Sept	81.5	81.5
Oct	82	75.8
Nov	64	74.0
Dec	76	

So, in this dataset we have some missing values of temperature in certain months like May and Septmenber. In this scenario, we can use moving average as a data imputation method where we calculate the averages of previous 2-3 or more records and impute those values in the missing rows.

This color represents, the cells were having missing values and we have imputed the moving average values in those cells.

Helper Column

67445	Total Blank	0
64950	Total Count	2240
61839	% of Blank	0.00%
52247	Mean	52247.24866
72335	Median	51741.5
36864	STDEV	25037.79717
66503	CV	47.9%

"Helper Column" -> Hide Them