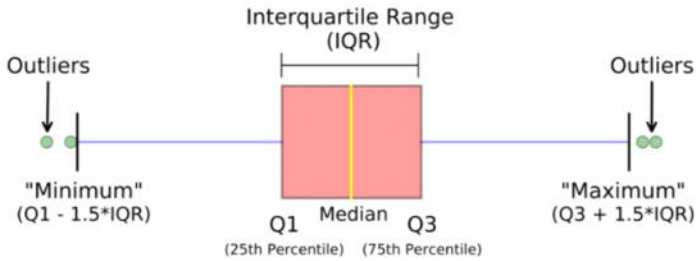
Handling Outliers & Variability + Data Analysis - I

🎯 Learning Goals:

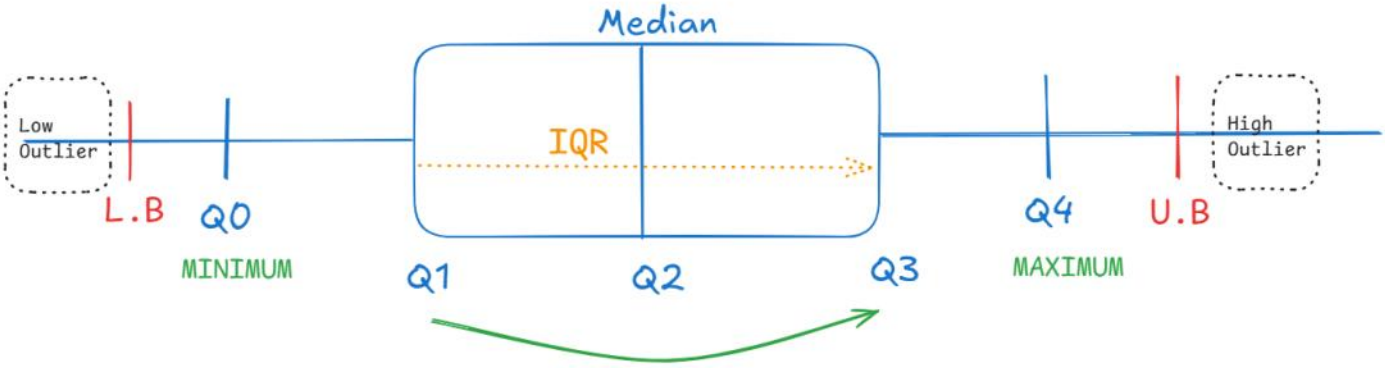
- Understand the basics of feature engineering.
- Learn pivot tables and their powerful data summarization capabilities.
- Explore advanced pivot table features like filters, grouping, slicers, and GETPIVOTDATA.
- Create charts and analyze data visually in Excel.



Handling Outlier

QUARTILE - Q

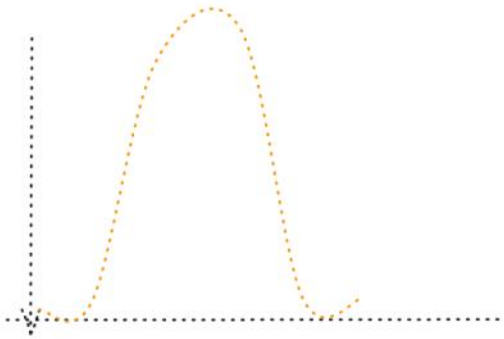
Q2 is close to median ~



Q3 - Q1 = IQR [Interquartile Range]

Lower Bound - Q1 - (1.5 times IQR)

Upper Bound - Q3 + (1.5 times IQR)



any value < L.B [Low outliers]

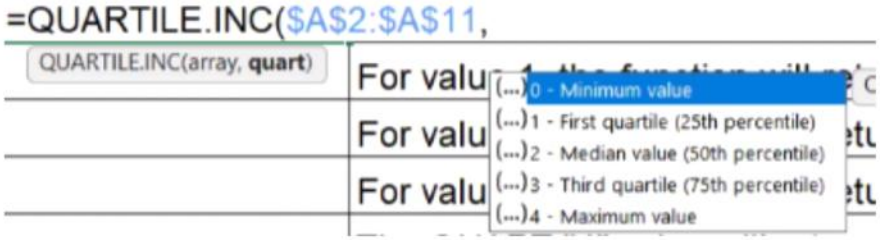
any value > U.B [High outliers]

INCLUSIVE ->

0 - 10 -- 20 -- 30 -- ..... -- 100

[0,100] Inclusive. >=0 AND <=100.

EXCLUSIVE -> [0,100] Exclusive >0 AND <100



SALES REVENUE
45000
23000
32000
11000
67000
89000
50000
17000
25000
57000

QUART VALUE	RESULT	
0	11000	=QUARTILE.INC($A$2:$A$11,0)
1	23500	=QUARTILE.INC($A$2:$A$11,1)
2	38500	=QUARTILE.INC($A$2:$A$11,2)
3	55250	=QUARTILE.INC($A$2:$A$11,3)
4	89000	=QUARTILE.INC($A$2:$A$11,4)

QUART VALUE	RESULT
0	11000
1	23500
2	38500
3	55250
4	89000
IQR	= D5-D3

IQR = Q3 - Q1.

QUART VALUE	RESULT
0	11000
1	23500
2	38500
3	55250
4	89000
IQR(Q3-Q1)	31750
Lower Bound	=D3-(1.5*D8)

LOWER BOUND = Q1 - (1.5 * IQR)

QUART VALUE	RESULT
0	11000
1	23500
2	38500
3	55250
4	89000
IQR(Q3-Q1)	31750
Lower Bound	-24125
Upper Bound	=D5+(1.5*D8)

UPPER BOUND = Q3 + (1.5 * IQR)

There is no outlier present in the dataset.

SALES REVENUE
45000
23000
32000
11000
67000
89000
50000
17000
25000
57000

QUART VALUE	RESULT
0	11000
1	23500
2	38500
3	55250
4	89000
IQR(Q3-Q1)	31750
Lower Bound	-24125
Upper Bound	102875

HARMEAN | =IF(OR(B14<$F$8,B14>$F$9),"Outlier","Not a Outlier")

OR(logical1, [logical2], [logical3], ...)

	A	B	C	E	F
1	STUDENTS NAME	TEST SCORE	Identify Outliers	QUARTILE VALU	RESULT
2	Alberta Fabela	79	Not a Outlier	0	58
3	Betty scruggs	73	Not a Outlier	1	71.75
4	Debbie Orone	94	Not a Outlier	2	83.5
5	Elvin Catron	58	Not a Outlier	3	94.25
6	Era Gassner	73	Not a Outlier	4	194
7	Grisede Mariscal	84	Not a Outlier	IQR	22.5
8	Gustano Primm	58	Not a Outlier	Lower Bound	38
9	Jacinda Moffett	68	Not a Outlier	Upper Bound	128
10	Juan Luker	95	Not a Outlier		
11	Karisa Oquwndo	155	Outlier		
12	Lacresha Whitty	61	Not a Outlier		
13	Laticia Crimi	75	Not a Outlier		
14	Loralee Widell	194	B14>$F$9),		
15	Octavio Ricci	64	Not a Outlier		
16	Ricky Cheshire	91	Not a Outlier		
17	Season Viers	94	Not a Outlier		
18	Shan texeria	99	Not a Outlier		
19	Sylvester Willbourn	85	Not a Outlier		
20	Taylor Pogue	83	Not a Outlier		
21	Yee Bordeau	111	Not a Outlier		

=IF(OR(B14<$F$8,B14>$F$9),"Outlier","Not a Outlier")

2 filters

STUDENTS NAME	TEST SCORE	Identify Outliers
Alberta Fabela	79	Not a Outlier
Betty scruggs	3	Outlier
Debbie Orone	94	Not a Outlier
Elvin Catron	58	Not a Outlier
Era Gassner	73	Not a Outlier
Grisede Mariscal	84	Not a Outlier
Gustano Primm	58	Not a Outlier
Jacinda Moffett	68	Not a Outlier
Juan Luker	95	Not a Outlier
Karisa Oquwndo	155	Outlier
Lacresha Whitty	61	Not a Outlier
Laticia Crimi	75	Not a Outlier
Loralee Widell	194	Outlier
Octavio Ricci	64	Not a Outlier
Ricky Cheshire	91	Not a Outlier
Season Viers	94	Not a Outlier
Shan texeria	99	Not a Outlier
Sylvester Willbourn	85	Not a Outlier
Taylor Pogue	83	Not a Outlier
Yee Bordeau	111	Not a Outlier

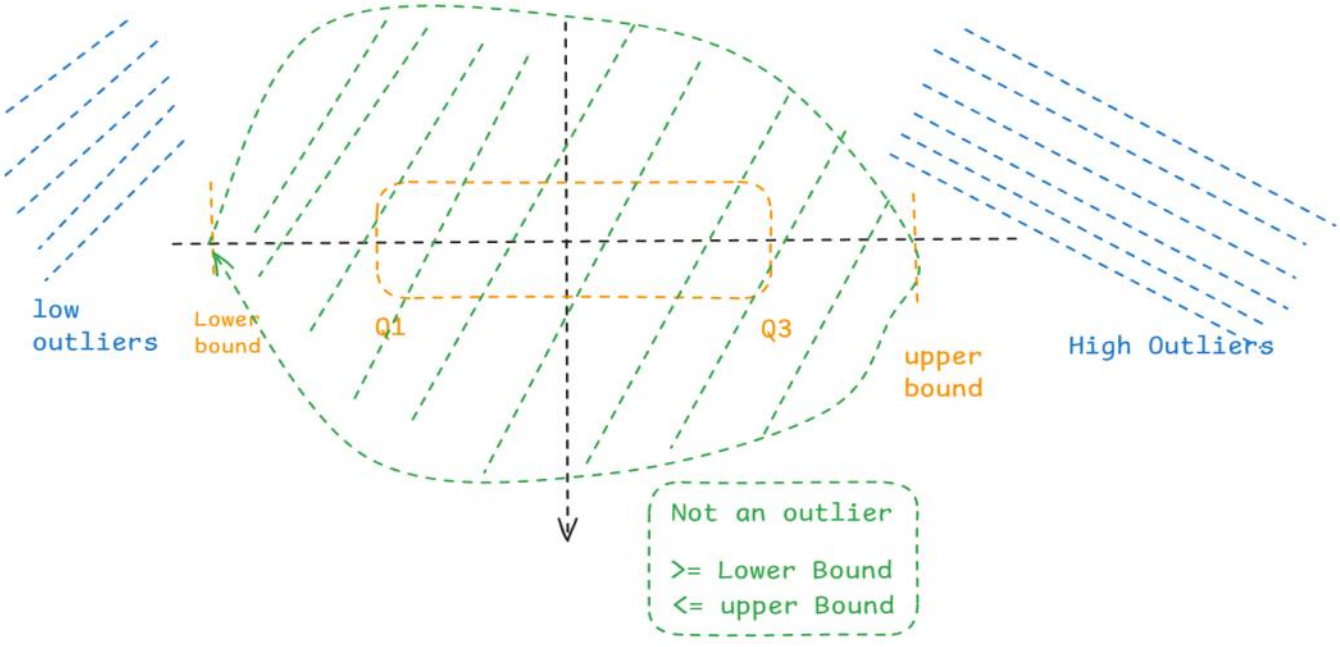
QUARTILE VALU	RESULT
0	3
1	67
2	83.5
3	94.25
4	194
IQR	27.25
Lower Bound	26.125
Upper Bound	135

Filter it down to High Outliers or Low Outliers. [Hint : IFS]

HARMEAN

```
=IFS(B2<$G$8,"Low Outlier",B2>$G$9,"High Outlier",OR(B2>=$G$8,B2<=$G$9),"Not an Outlier")
```

	A	B	C	D	E	F	G
1	STUDENTS NAME	TEST SCORE	Identify Outliers	Finding High/Low Outliers		QUARTILE VALU	RESULT
2	Alberta Fabela	79	Not a Outlier	B2<=$G$9),"Not an Outlier")		0	3
3	Betty scruggs	3	Outlier	Low Outlier		1	67
4	Debbie Orone	94	Not a Outlier	Not an Outlier		2	83.5
5	Elvin Catron	58	Not a Outlier	Not an Outlier		3	94.25
6	Era Gassner	73	Not a Outlier	Not an Outlier		4	194
7	Grisede Mariscal	84	Not a Outlier	Not an Outlier		IQR	27.25
8	Gustano Primm	58	Not a Outlier	Not an Outlier		Lower Bound	26.125
9	Jacinda Moffett	68	Not a Outlier	Not an Outlier		Upper Bound	135
10	Juan Luker	95	Not a Outlier	Not an Outlier			
11	Karisa Oquwndo	155	Outlier	High Outlier			
12	Lacresha Whitty	61	Not a Outlier	Not an Outlier			
13	Laticia Crimi	75	Not a Outlier	Not an Outlier			
14	Loralee Widell	194	Outlier	High Outlier			
15	Octavio Ricci	64	Not a Outlier	Not an Outlier			
16	Ricky Cheshire	91	Not a Outlier	Not an Outlier			
17	Season Viers	94	Not a Outlier	Not an Outlier			
18	Shan texeria	99	Not a Outlier	Not an Outlier			
19	Sylvester Willbourn	85	Not a Outlier	Not an Outlier			
20	Taylor Pogue	83	Not a Outlier	Not an Outlier			
21	Yee Bordeau	111	Not a Outlier	Not an Outlier			

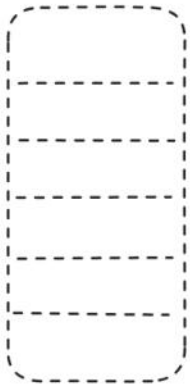


AND

```
=IF(OR(B2<$G$8,B2>$G$9),"Outlier","Not a Outlier")
```

```
=IFS(B2<$G$8,"Low Outlier",
     B2>$G$9,"High Outlier",
     OR(B2>=$G$8,B2<=$G$9),"Not an Outlier")
```

AND



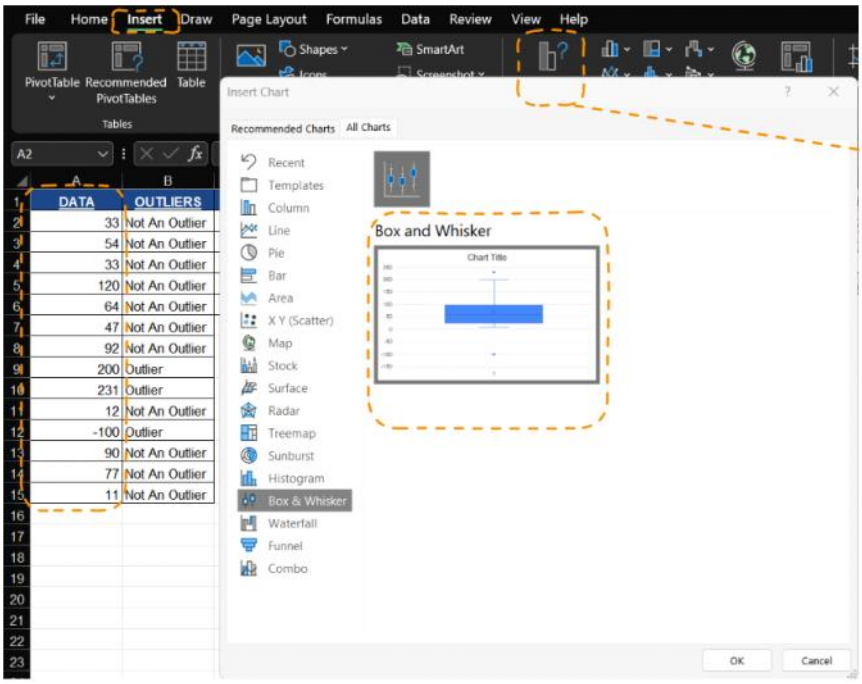
IF ELSE Ladder

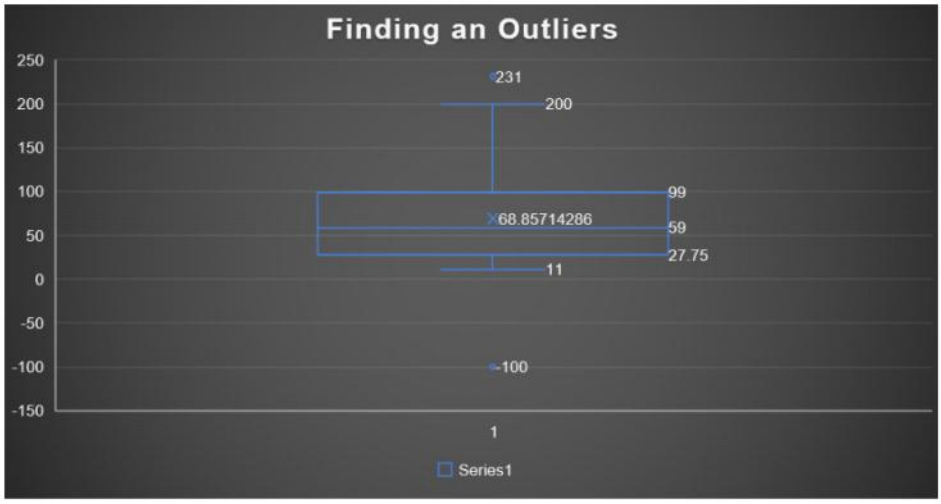
DATA	OUTLIERS	QUARTILE	RESULTS	IQR (Q3-Q1)	UPPER BOUND	LOWER BOUND
33	Not An Outlier	0	-100	58.5	179.25	-54.75
54	Not An Outlier	1	33			
33	Not An Outlier	2	59			
120	Not An Outlier	3	91.5			
64	Not An Outlier	4	231			
47	Not An Outlier					
92	Not An Outlier					
200	Outlier					
231	Outlier					
12	Not An Outlier					
-100	Outlier					
90	Not An Outlier					
77	Not An Outlier					
11	Not An Outlier					

IQR= Inter Quartile Range
Upper bound= Q3+1.5*IQR
Upper bound= Q1-1.5*IQR

lower than the "Lower bound", then those data will be considered as outliers.

DATA	OUTLIERS	QUARTILE	RESULTS	IQR (Q3-Q1)	UPPER BOUND	LOWER BOUND
33	=IF(OR(A2<$G$2,A2>$F$2),"(	0	=QUARTILE.INC($A$2:$A$15,	=D5-D3	=D5 + 1.5*E2	=D3-(1.5*E2)
54	=IF(OR(A3<$G$2,A3>$F$2),"(	1	=QUARTILE.INC($A$2:$A$15,			
33	=IF(OR(A4<$G$2,A4>$F$2),"(	2	=QUARTILE.INC($A$2:$A$15,			
120	=IF(OR(A5<$G$2,A5>$F$2),"(	3	=QUARTILE.INC($A$2:$A$15,			
64	=IF(OR(A6<$G$2,A6>$F$2),"(	4	=QUARTILE.INC($A$2:$A$15,			
47	=IF(OR(A7<$G$2,A7>$F$2),"(					
92	=IF(OR(A8<$G$2,A8>$F$2),"(					
200	=IF(OR(A9<$G$2,A9>$F$2),"(					
231	=IF(OR(A10<$G$2,A10>$F$2)					
12	=IF(OR(A11<$G$2,A11>$F$2)					
-100	=IF(OR(A12<$G$2,A12>$F$2)					
90	=IF(OR(A13<$G$2,A13>$F$2)					
77	=IF(OR(A14<$G$2,A14>$F$2)					
11	=IF(OR(A15<$G$2,A15>$F$2)					


File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Help
PivotTable
Recommended PivotTables
Table
Tables
Shapes
SmartArt
Insert Chart
Recommended Charts
All Charts
Recent
Templates
Column
Line
Pie
Bar
Area
X Y (Scatter)
Map
Stock
Surface
Radar
Treemap
Sunburst
Histogram
Box & Whisker
Waterfall
Funnel
Combo
Box and Whisker
Chart Title
OK
Cancel
DATA
OUTLIERS
33 Not An Outlier
54 Not An Outlier
33 Not An Outlier
120 Not An Outlier
64 Not An Outlier
47 Not An Outlier
92 Not An Outlier
200 Outlier
231 Outlier
12 Not An Outlier
-100 Outlier
90 Not An Outlier
77 Not An Outlier
11 Not An Outlier
ALL Charts


Finding an Outliers
250
200
150
100
50
0
-50
-100
-150
231
200
99
68.85714286
59
27.75
11
-100
1
Series1

Data Analysis - I

What is Feature Engineering?

- It is a process of selecting, transforming and creating new features on a particular table.

DOB --> Age

location Time

Delhi , 11:00AM
Kolkata , 10:05PM

Location

Delhi
Kolkata

Time

11:00 AM
10:05 PM

AGE ranging from 0 - 100 is also not a idea call to find an insights.

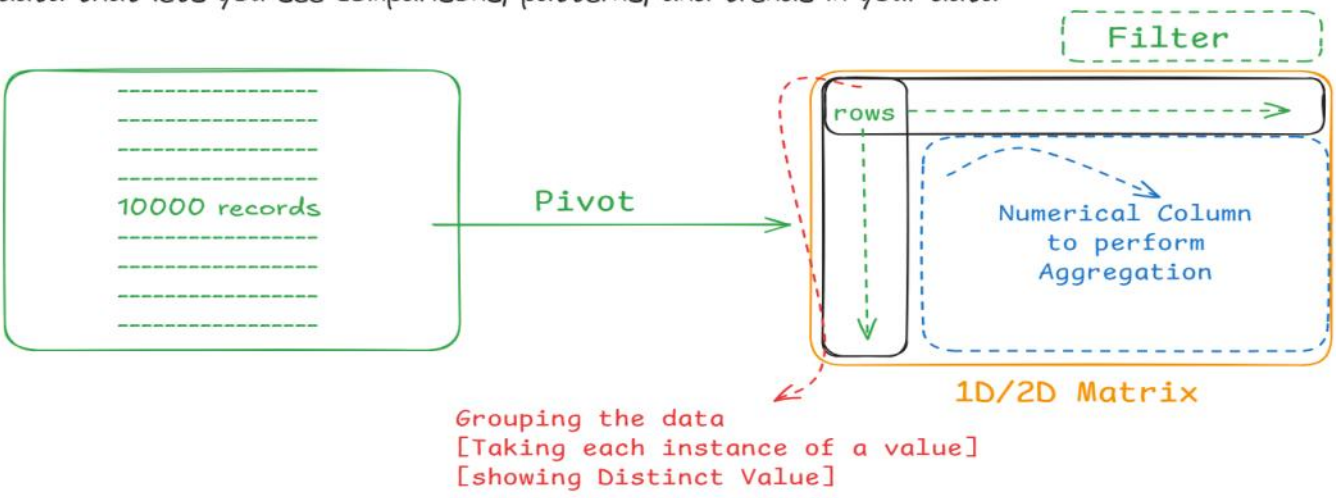
0-10 -> Kids
11-18 -> Teens
18-35 -> Adults
35-60 -> Adult Pro Max
60+ -> Seniors

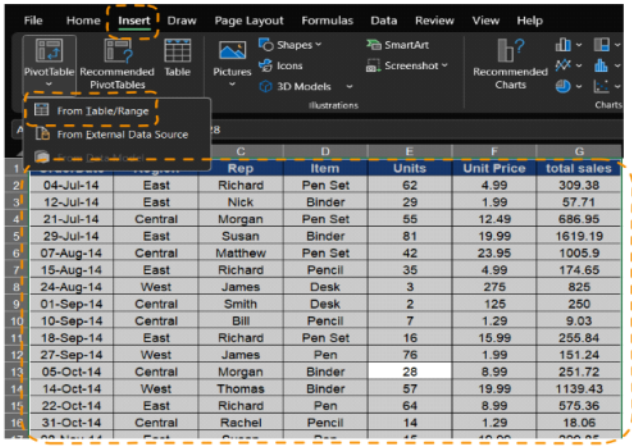
Sales Transaction Tables

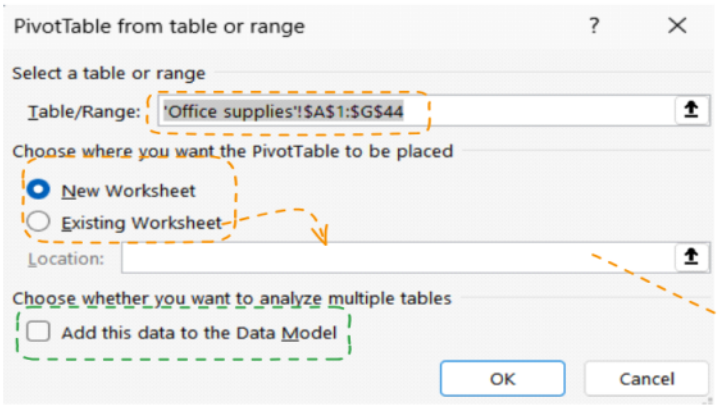
Product Price	Product Cost	Order Qty	Sales = Product Price * Order Qty	Expenses = Product Cost * Order Qty	Profit = Sales-Expenses

What is Pivot Table? "Summarization"

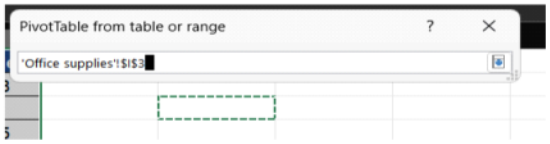
Pivot table is like a summary table of a larger data to get insights on few aspects, to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data

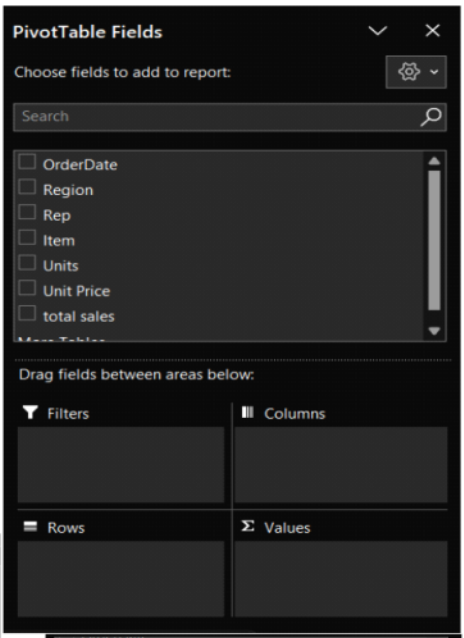


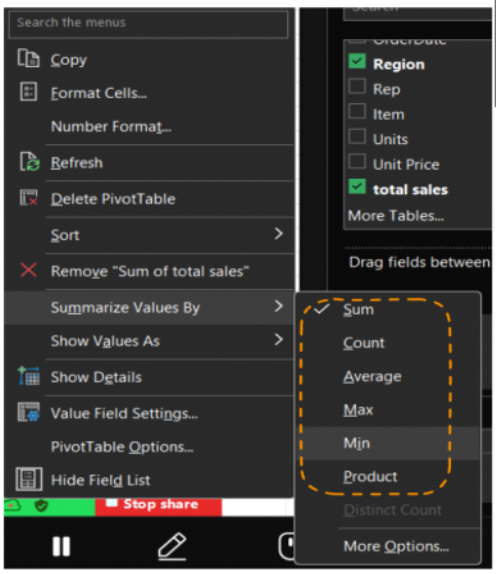


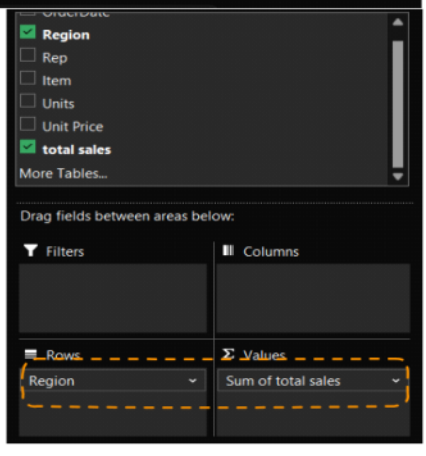


Provide the location of the cell where you want to create a pivot









Row Labels	Sum of total sales
Central	11139.07
East	6002.09
West	2486.72
Grand Total	**19627.88**

Find the Total Sales from Each Region.

region has duplicate values.
-> Distinct Value
-> Group By

DISTINCT VS UNIQUE

DISTINCT - 5
one or more Occurrence
Final show at least one

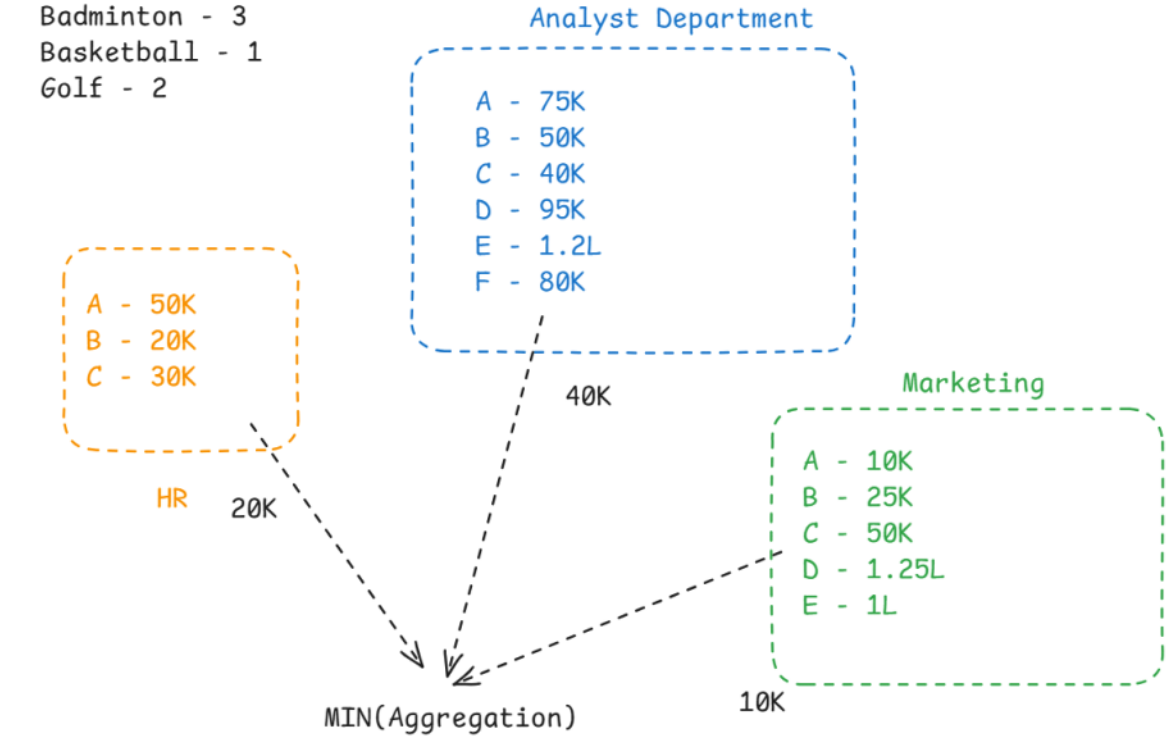
UNIQUE - 1
(only 1 occurrence)

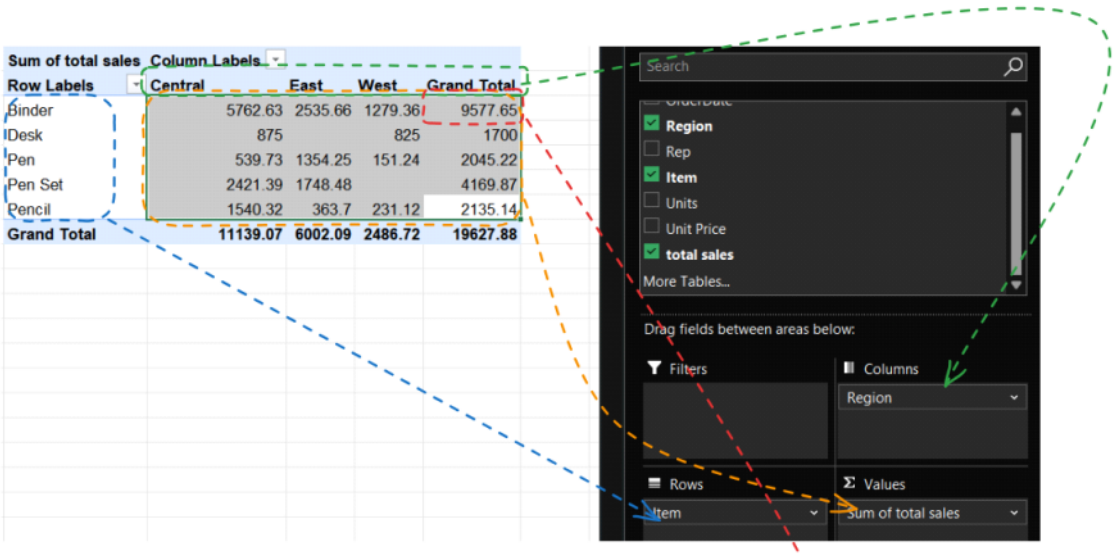
Football Golf Cricket Football Badminton Football Cricket Football Badminton Football Cricket Golf Cricket Football Football BasketBall Badminton

Cricket
Football
Badminton
Basketball
Golf

Basketball

Cricket - 4
Football - 7
Badminton - 3
Basketball - 1
Golf - 2





OrderDate	Region	Rep	Item	Units	Unit Price	total sales
31-05-2015	Central	Bill	Binder	80	8.99	719.2
01-02-2015	Central	Smith	Binder	87	15	1305
23-01-2015	Central	Matthew	Binder	50	19.99	999.5
15-01-2015	Central	Bill	Binder	46	8.99	413.54
21-12-2014	Central	Rachel	Binder	28	4.99	139.72
05-10-2014	Central	Morgan	Binder	28	8.99	251.72
04-12-2014	Central	Alex	Binder	94	19.99	1879.06
17-11-2014	Central	Alex	Binder	11	4.99	54.89
08-06-2015	East	Richard	Binder	60	8.99	539.4
12-07-2014	East	Nick	Binder	29	1.99	57.71
29-07-2014	East	Susan	Binder	81	19.99	1619.19
01-04-2015	East	Richard	Binder	60	4.99	299.4
18-02-2015	East	Richard	Binder	4	4.99	19.96
07-03-2015	West	James	Binder	7	19.99	139.93
14-10-2014	West	Thomas	Binder	57	19.99	1139.43

Double click to see granular detail.