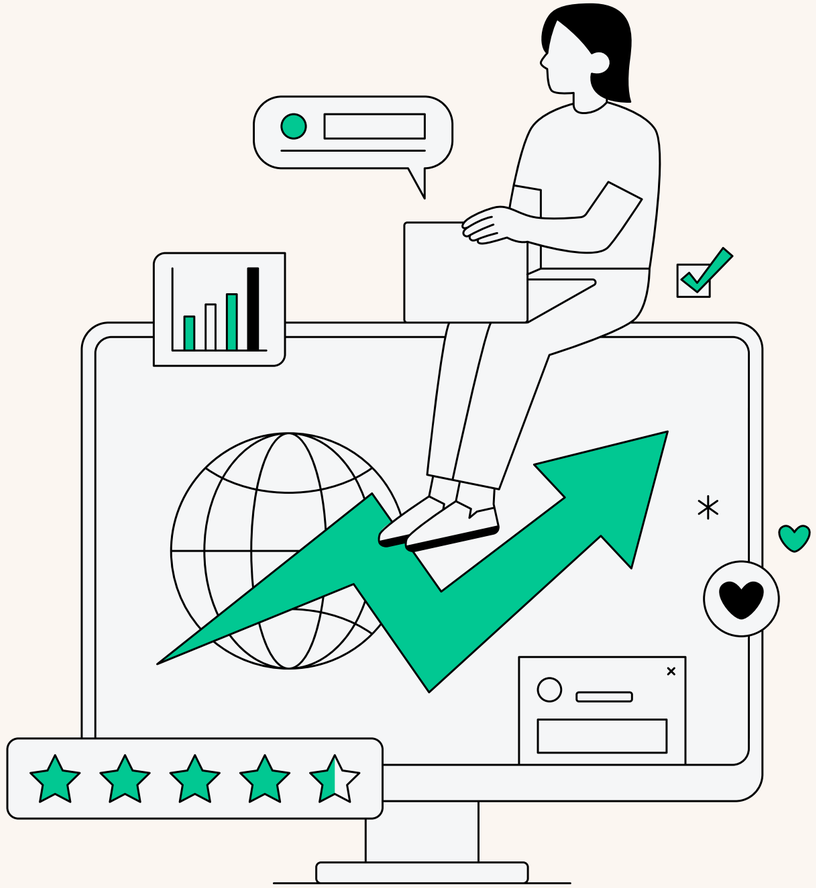
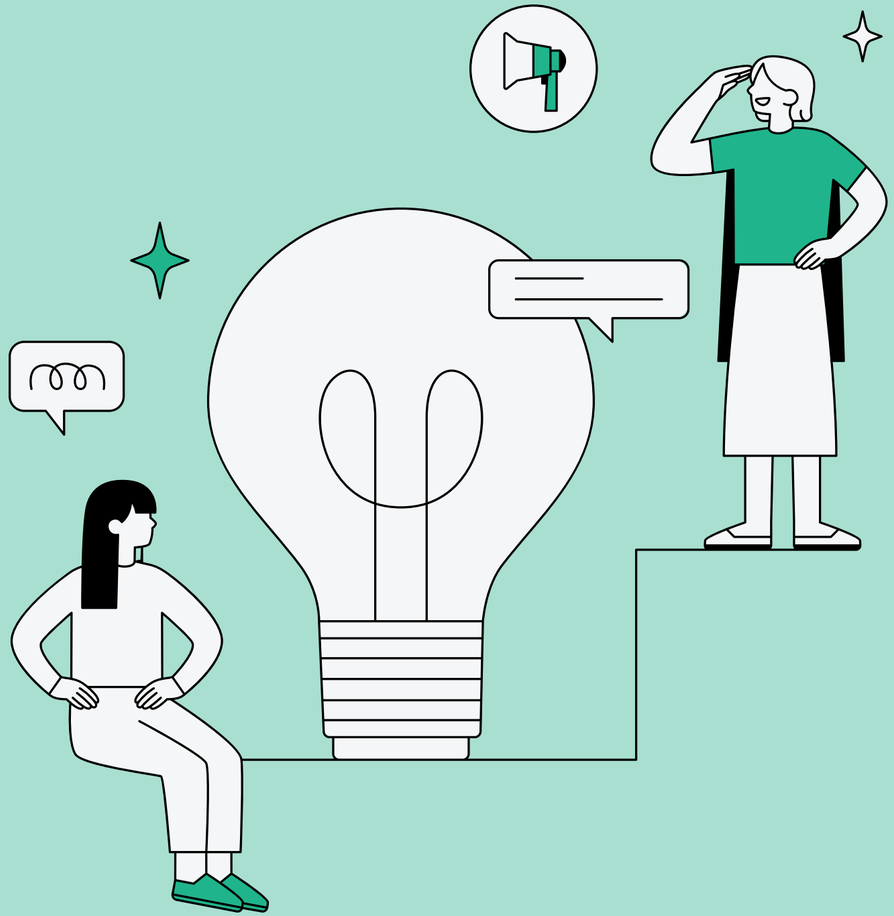
Presented by Azam
02-10-2024

Myntra Dataset Analysis

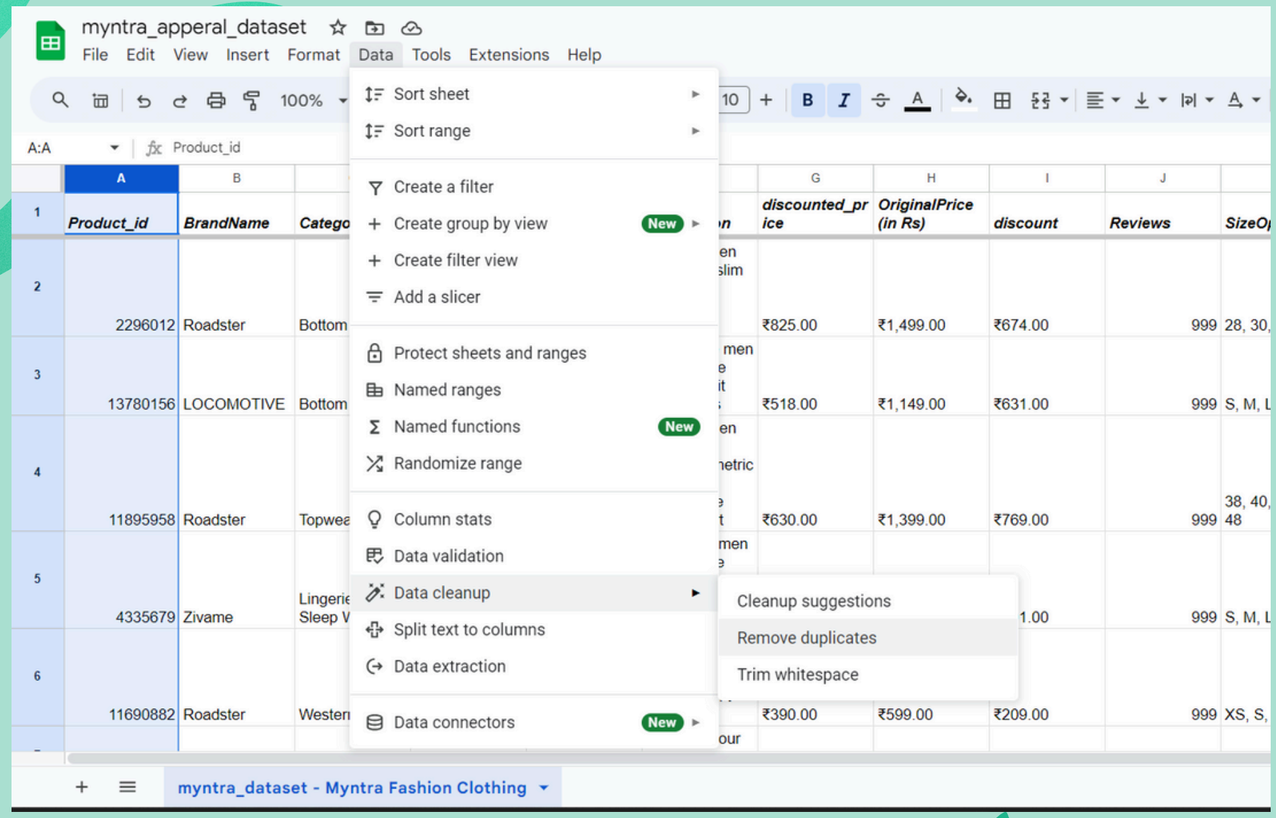


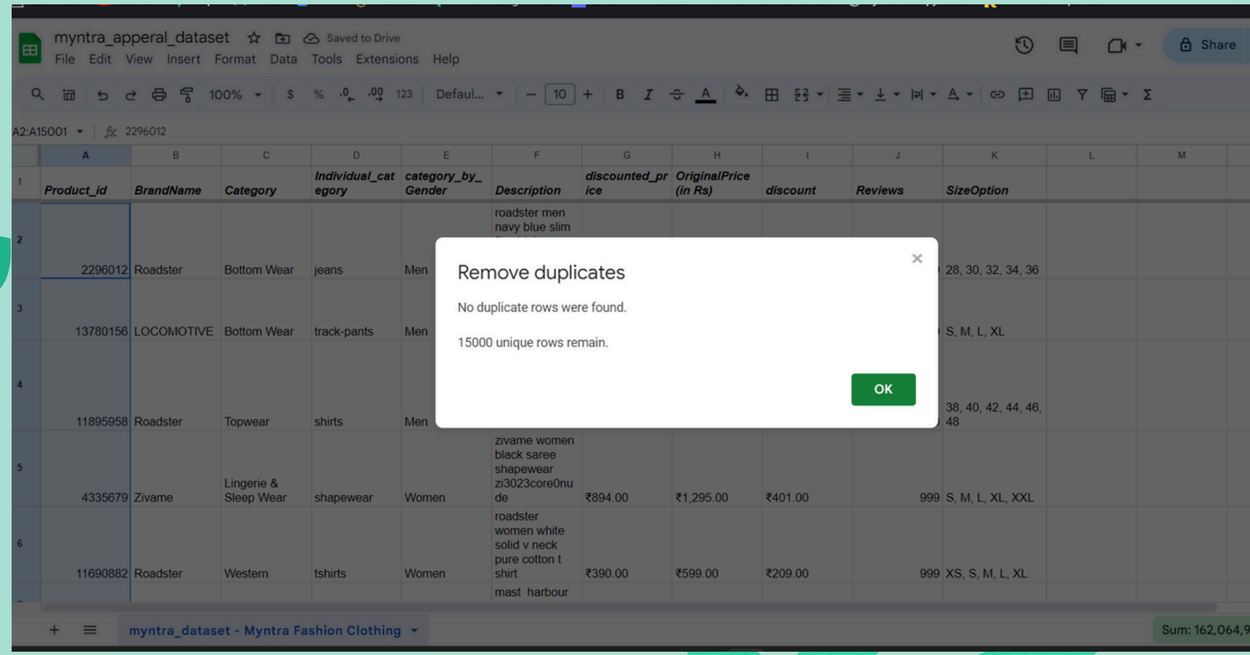
Data Cleaning

Data cleaning is essential for accurate analysis and good decision-making. It involves fixing errors, inconsistencies, and inaccuracies in data, such as typos, wrong formats, and missing values. Clean data improves data quality and allows for reliable insights, leading to better business outcomes.



Finding Duplicates

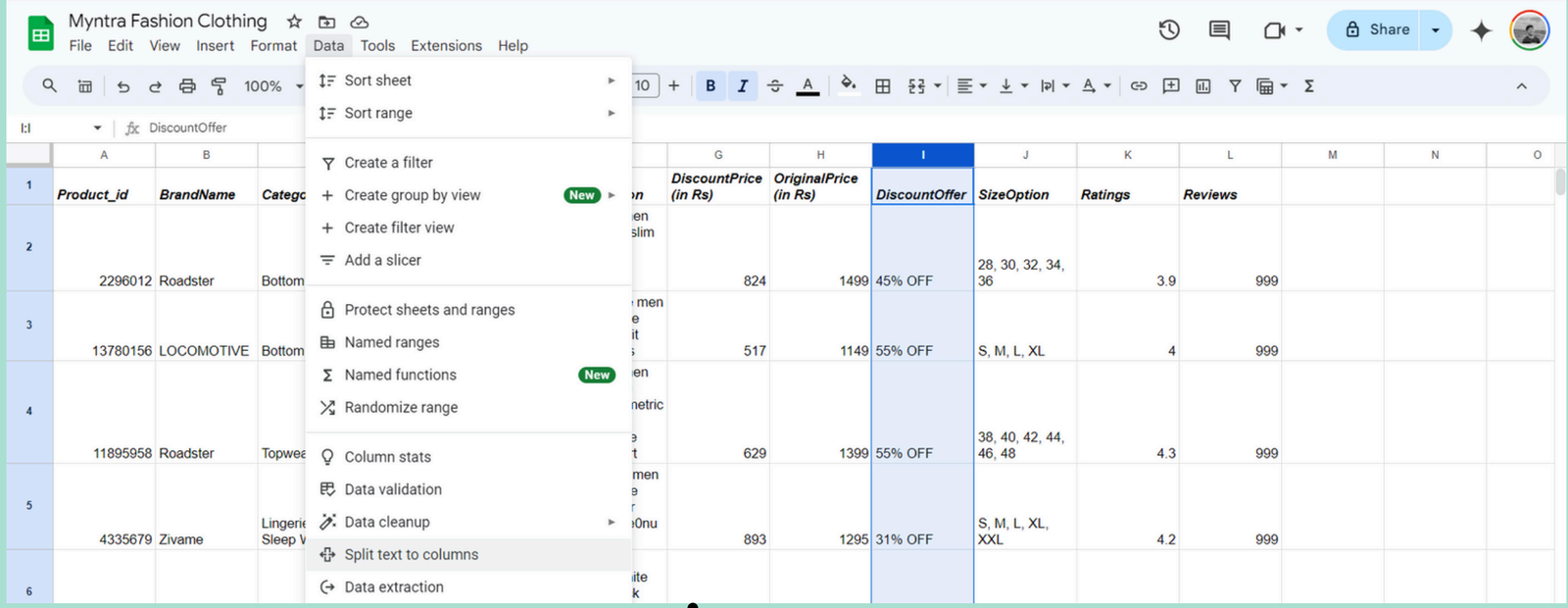

myntra_apperal_dataset
File Edit View Insert Format Data Tools Extensions Help
Sort sheet
Sort range
Create a filter
Create group by view
New
Create filter view
Add a slicer
Protect sheets and ranges
Named ranges
Named functions
New
Randomize range
Column stats
Data validation
Data cleanup
Split text to columns
Data extraction
Data connectors
New
Cleanup suggestions
Remove duplicates
Trim whitespace
Product_id
BrandName
discounted_price
OriginalPrice (in Rs)
discount
Reviews
2296012 Roadster ₹825.00 ₹1,499.00 ₹674.00 999
13780156 LOCOMOTIVE ₹518.00 ₹1,149.00 ₹631.00 999
11895958 Roadster ₹630.00 ₹1,399.00 ₹769.00 999
4335679 Zivame 999
11690882 Roadster ₹390.00 ₹599.00 ₹209.00 999
myntra_dataset - Myntra Fashion Clothing

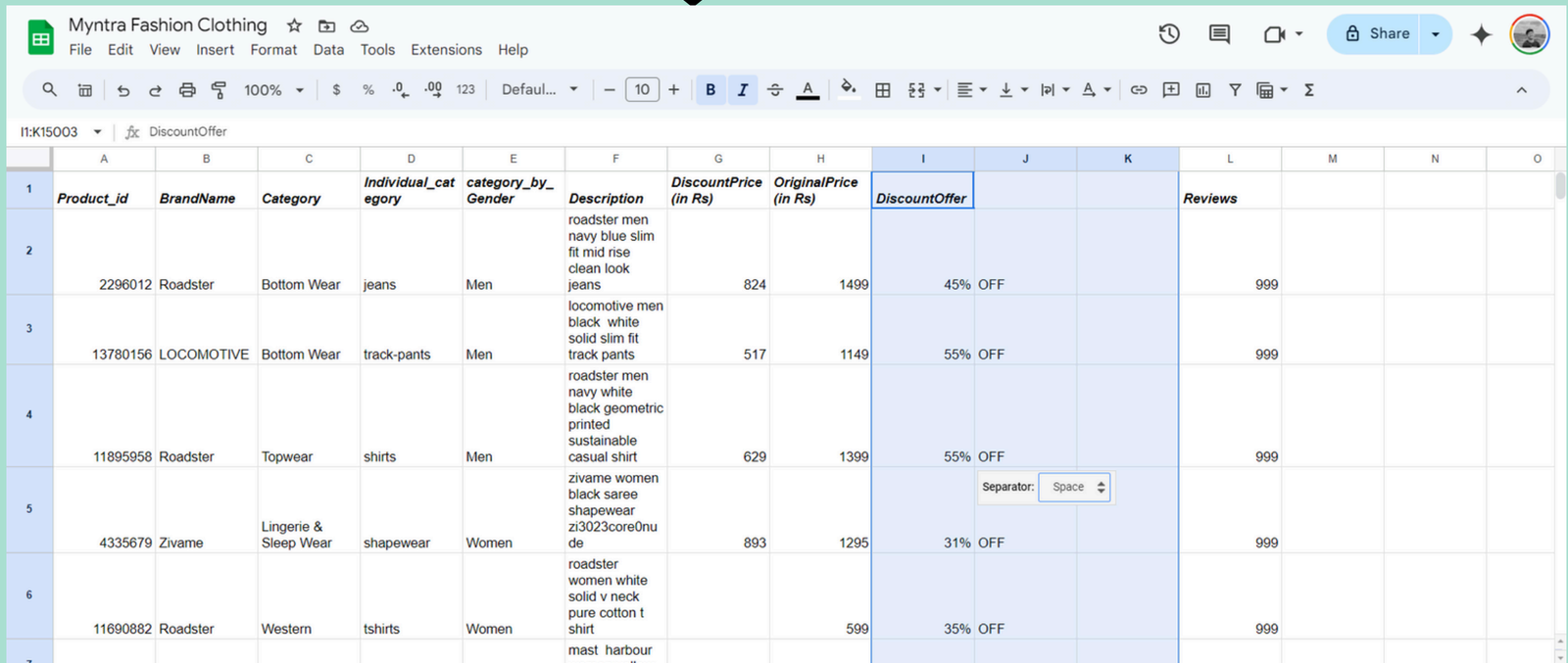

myntra_apperal_dataset
Saved to Drive
File Edit View Insert Format Data Tools Extensions Help
Share
Remove duplicates
No duplicate rows were found.
15000 unique rows remain.
OK
Product_id
BrandName
Category
Individual_category
category_by_Gender
Description
discounted_price
OriginalPrice (in Rs)
discount
Reviews
SizeOption
2296012 Roadster Bottom Wear jeans Men roadster men navy blue slim 28, 30, 32, 34, 36
13780156 LOCOMOTIVE Bottom Wear track-pants Men S, M, L, XL
11895958 Roadster Topwear shirts Men 38, 40, 42, 44, 46, 48
4335679 Zivame Lingerie & Sleep Wear shapewear Women zivame women black saree shapewear zi3023core0nude ₹894.00 ₹1,295.00 ₹401.00 999 S, M, L, XL, XXL
11690882 Roadster Western tshirts Women roadster women white solid v neck pure cotton t shirt ₹390.00 ₹599.00 ₹209.00 999 XS, S, M, L, XL
mast harbour
myntra_dataset - Myntra Fashion Clothing

Standardising Discount Offer Column

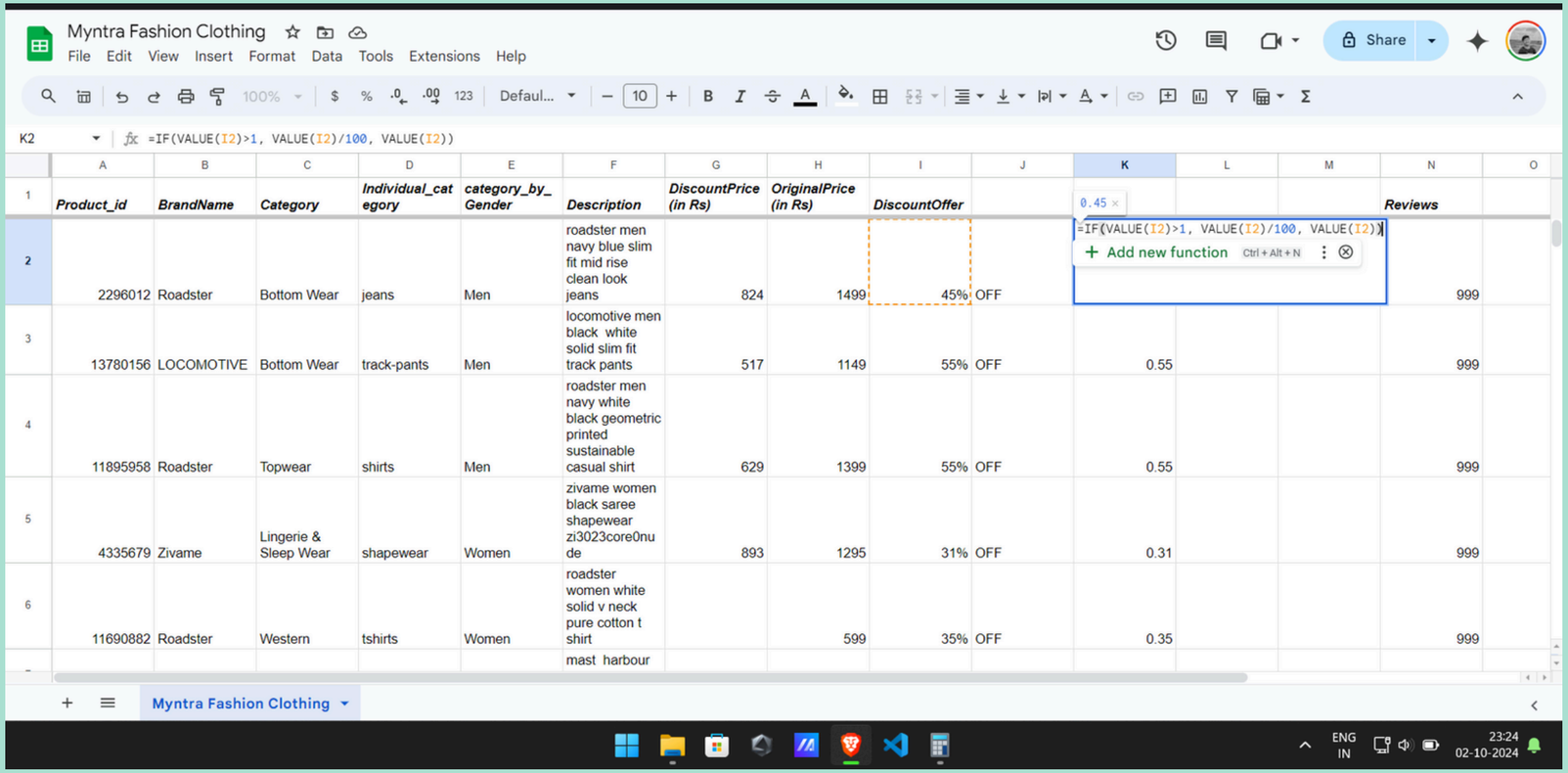
As the discount column contains both discount values in both % and Rs.

1. We split the column on space delimiter



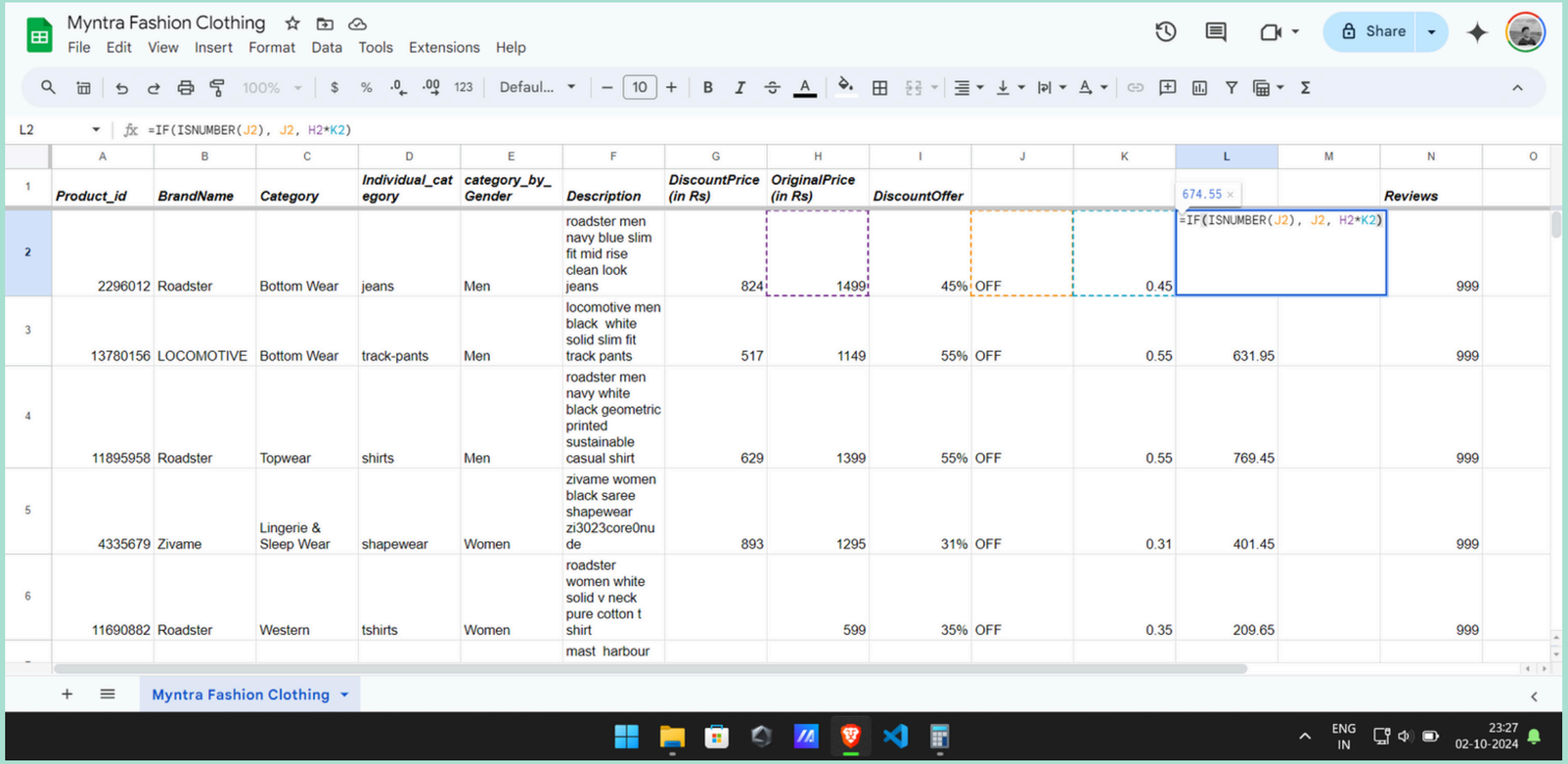


2. The 'Discount' column contains a mix of percentages and numbers, causing formatting issues. To standardize the data, a formula will convert any number greater than 1 to its decimal equivalent, ensuring all values represent percentages correctly.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	*Product_id*	*BrandName*	*Category*	*Individual_category*	*category_by_Gender*	*Description*	*DiscountPrice (in Rs)*	*OriginalPrice (in Rs)*	*DiscountOffer*					*Reviews*	
2	2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean look jeans	824	1499	45%	OFF				999	
3	13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit track pants	517	1149	55%	OFF	0.55			999	
4	11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric printed sustainable casual shirt	629	1399	55%	OFF	0.55			999	
5	4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0nude	893	1295	31%	OFF	0.31			999	
6	11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton t shirt		599	35%	OFF	0.35			999	
						mast harbour									

3. After ensuring that percentages are calculated correctly, now it’s time to convert and merge them all in the form of prices, which can be done by utilising the formula below



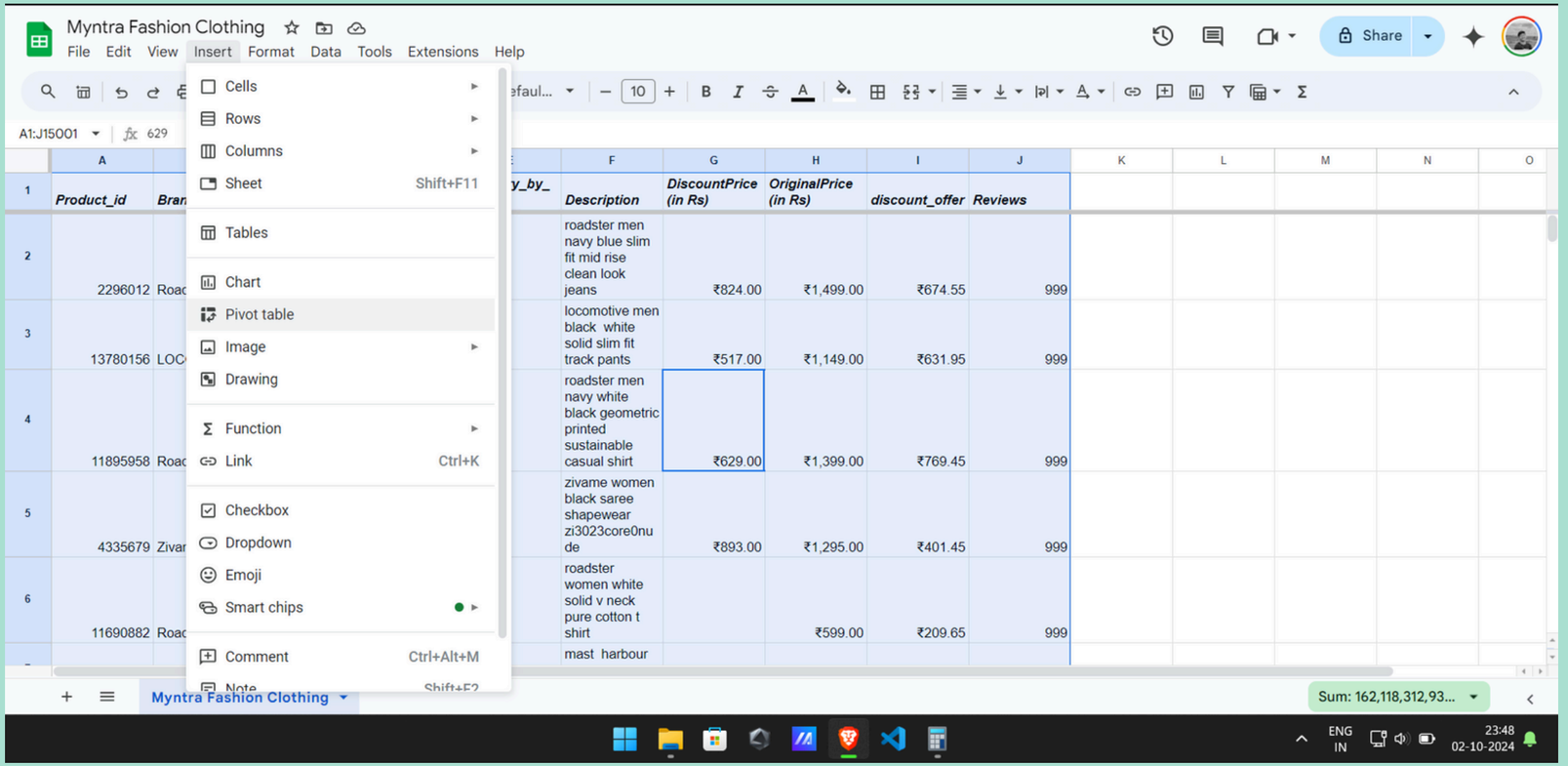
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	*Product_id*	*BrandName*	*Category*	*Individual_category*	*category_by_Gender*	*Description*	*DiscountPrice (in Rs)*	*OriginalPrice (in Rs)*	*DiscountOffer*					*Reviews*	
2	2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean look jeans	824	1499	45%	OFF	0.45	=IF(ISNUMBER(J2), J2, H2*K2)		999	
3	13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit track pants	517	1149	55%	OFF	0.55	631.95		999	
4	11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric printed sustainable casual shirt	629	1399	55%	OFF	0.55	769.45		999	
5	4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0nude	893	1295	31%	OFF	0.31	401.45		999	
6	11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton t shirt		599	35%	OFF	0.35	209.65		999	
						mast harbour									

Handling Missing Discounts: Imputation with Category Averages

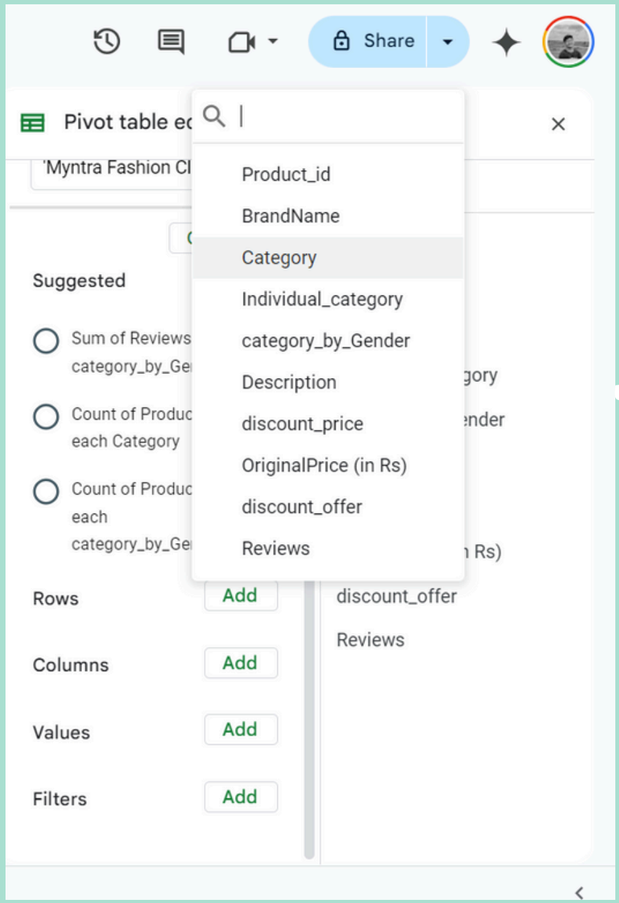
To find averages if we can utilise pivot tables to enhance our efficiency and not having to scratch our head using complex and nested formulas/functions

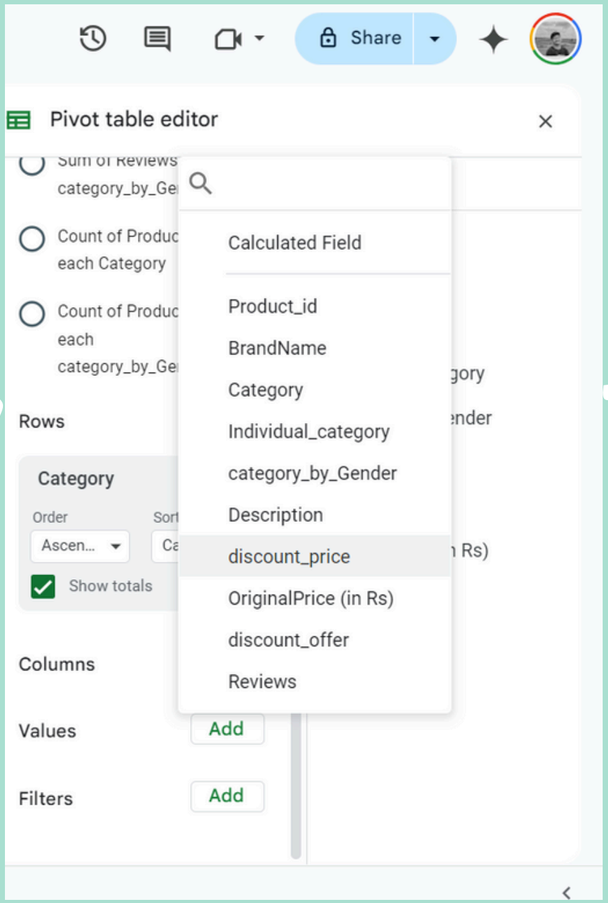
Steps to insert a pivot table:

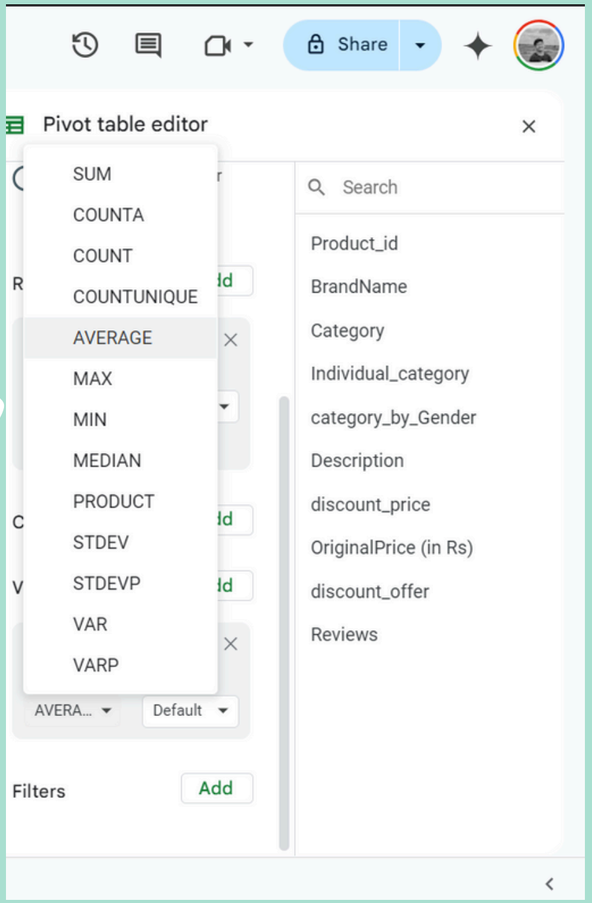
1. Select all data using Ctrl + A
2. Then go to Insert -> Pivot Table



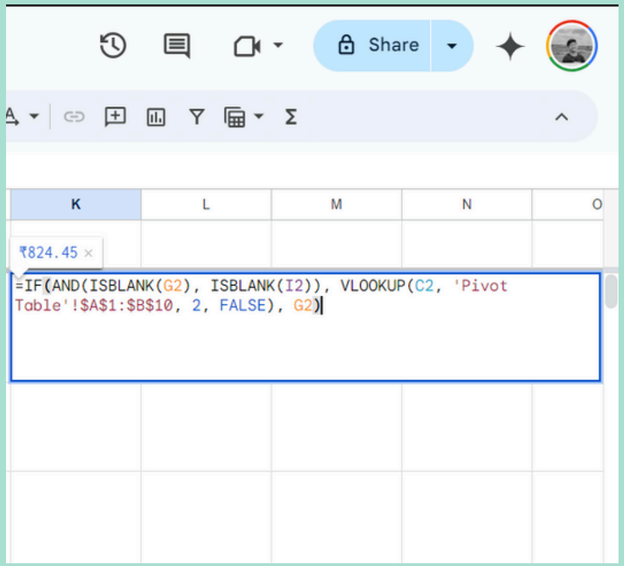
1. After opening the pivot table in new sheet, on the right hand side in the rows section click on add and select category.
2. Similarly, in the values section add discount_price
3. Select average/median instead of sum in the summarise by dropdown



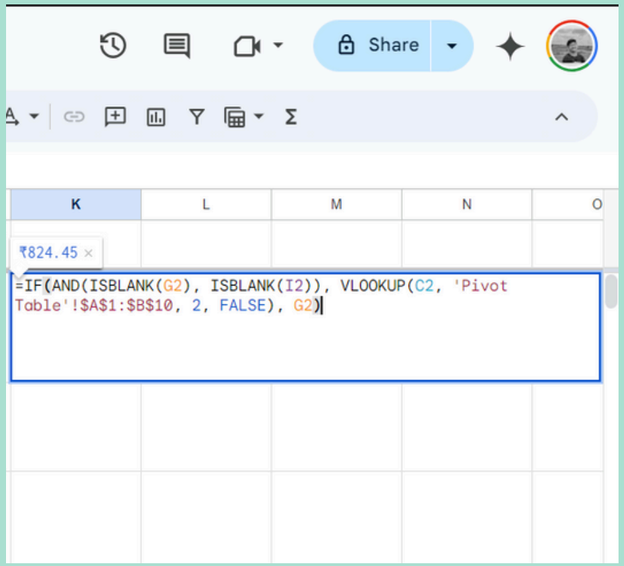




Now the average disount_price is calculated for each category.
Next, to utilise these values let’s use VLOOKUP inside an IF statement.



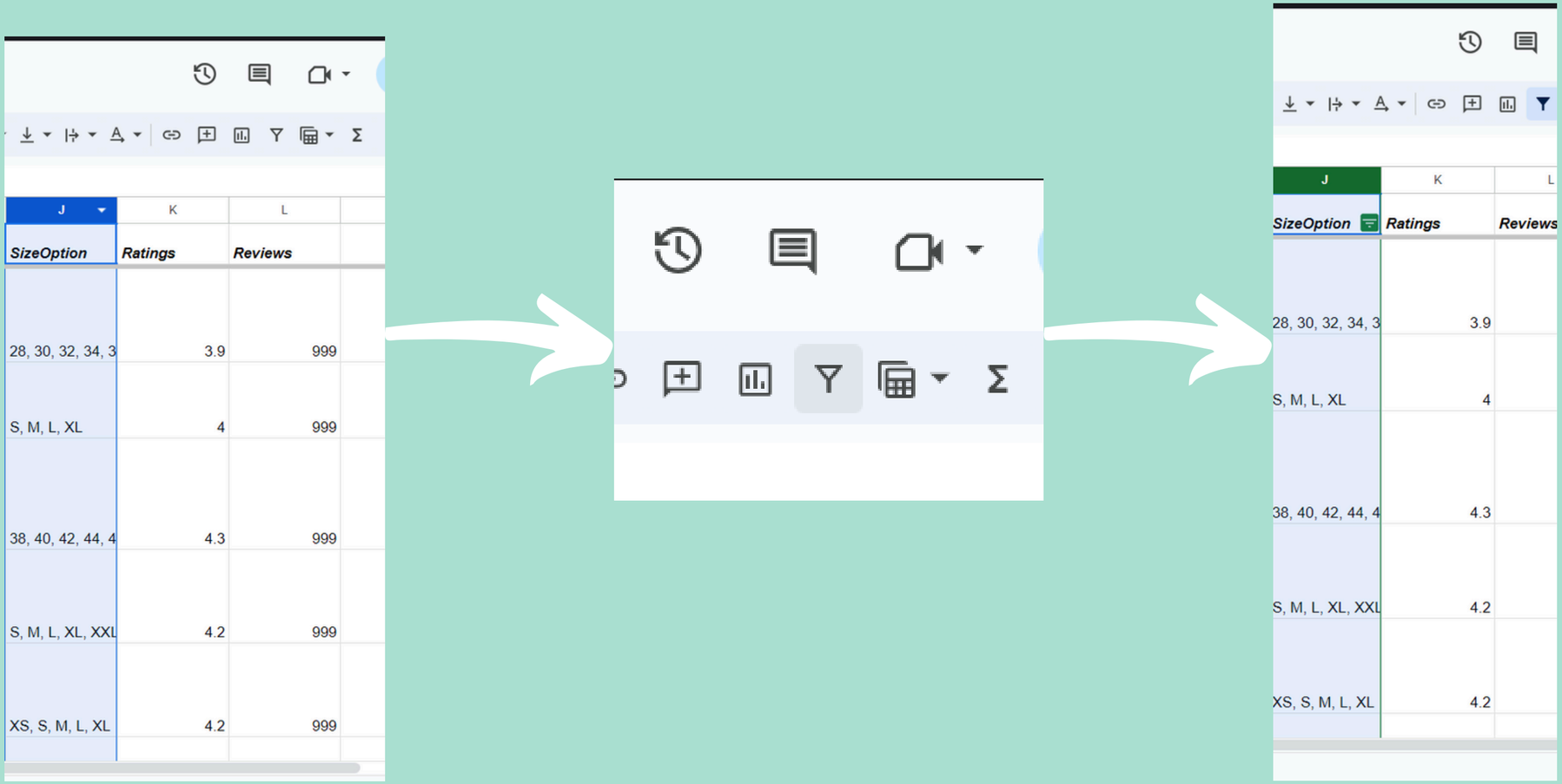
Now the average disount_price is calculated for each category.
Next, to utilise these values let’s use VLOOKUP inside an IF statement.



Replacing Null Values in 'SizeOption' Column with 'Not Available'

To replace all the null/blank values with N/A let’s first see if it contains any null values.

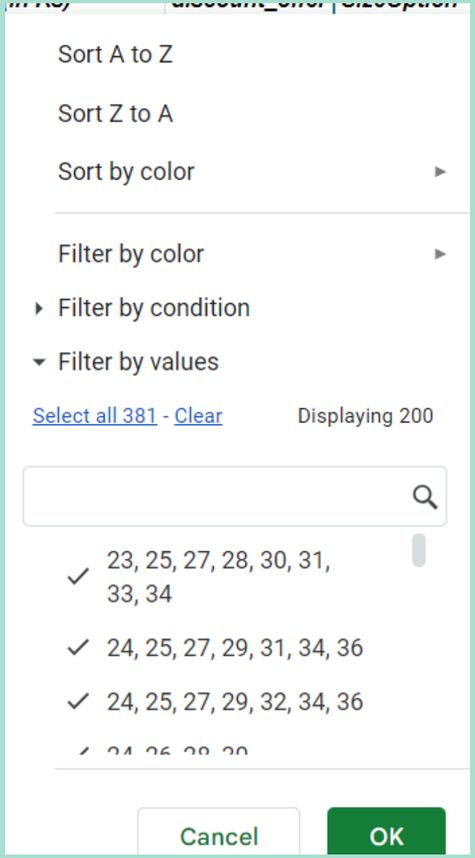
1. Select the complete column
2. Click the filter icon and check for (Blanks)



	J	K	L
	SizeOption	Ratings	Reviews
	28, 30, 32, 34, 3	3.9	999
	S, M, L, XL	4	999
	38, 40, 42, 44, 4	4.3	999
	S, M, L, XL, XXL	4.2	999
	XS, S, M, L, XL	4.2	999

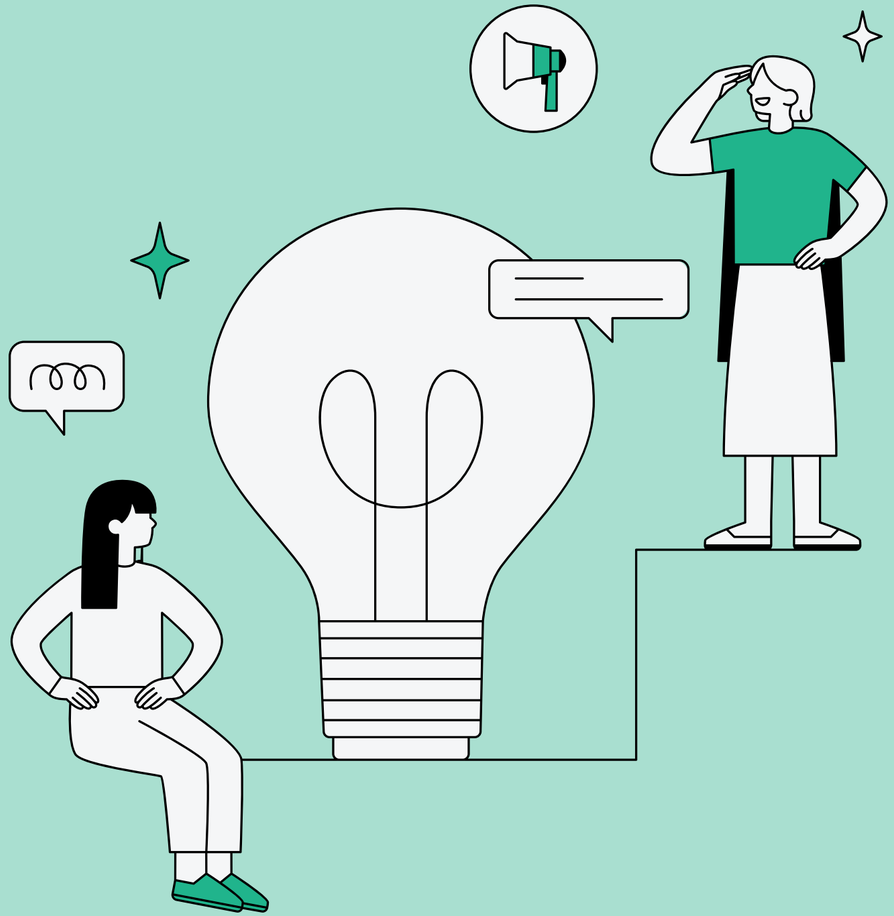
	J	K	L
	SizeOption	Ratings	Reviews
	28, 30, 32, 34, 3	3.9	
	S, M, L, XL	4	
	38, 40, 42, 44, 4	4.3	
	S, M, L, XL, XXL	4.2	
	XS, S, M, L, XL	4.2	

As we can see, we do not have any null values.



Data Analysis

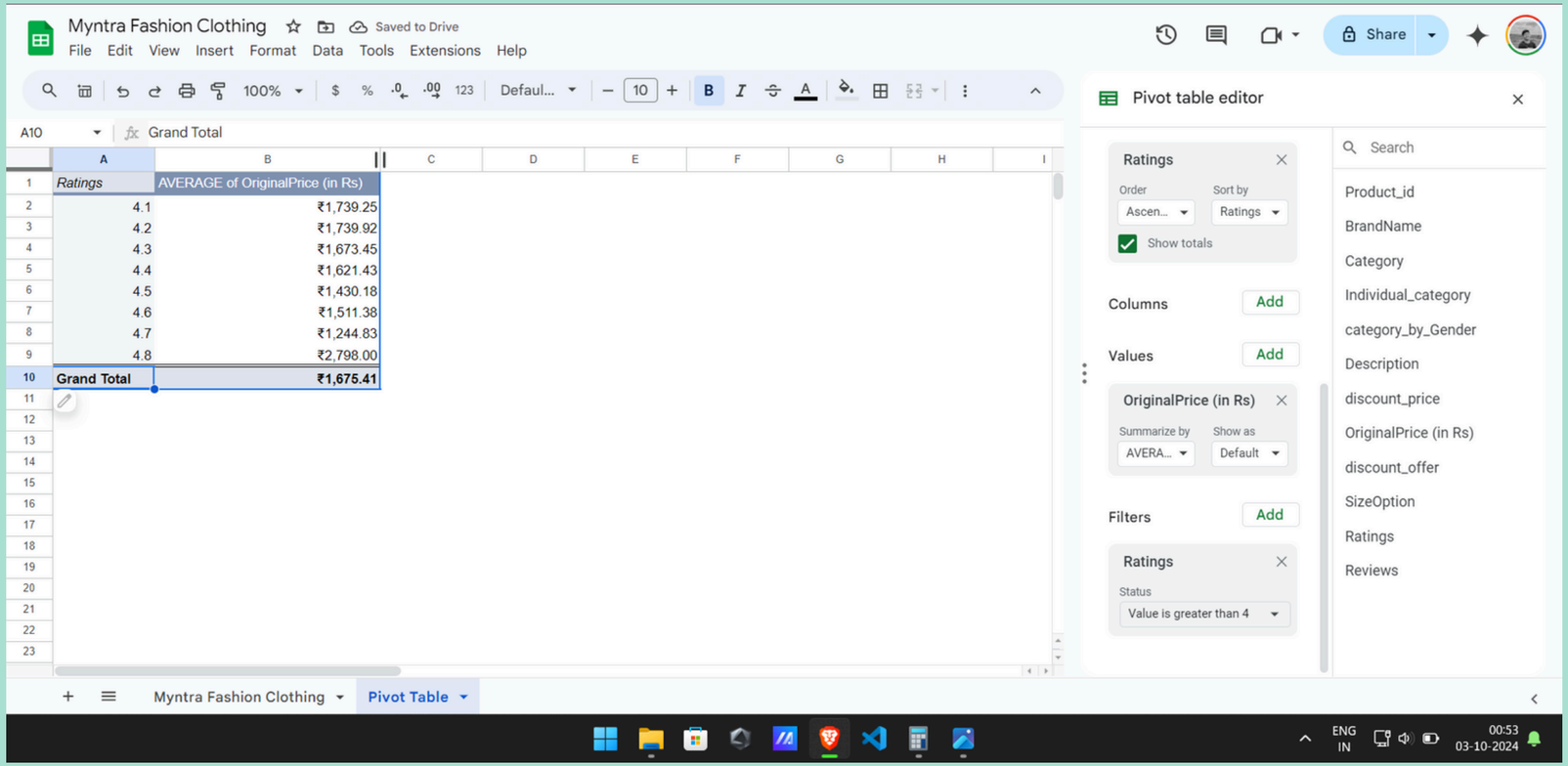
Data analysis involves examining raw data to uncover trends, patterns, and insights that can help with decision-making. It involves cleaning, transforming, and modeling data using various techniques. The goal is to extract meaningful information that can improve efficiency, solve problems, and drive better outcomes.



Overall average original price for products with ratings greater than 4.

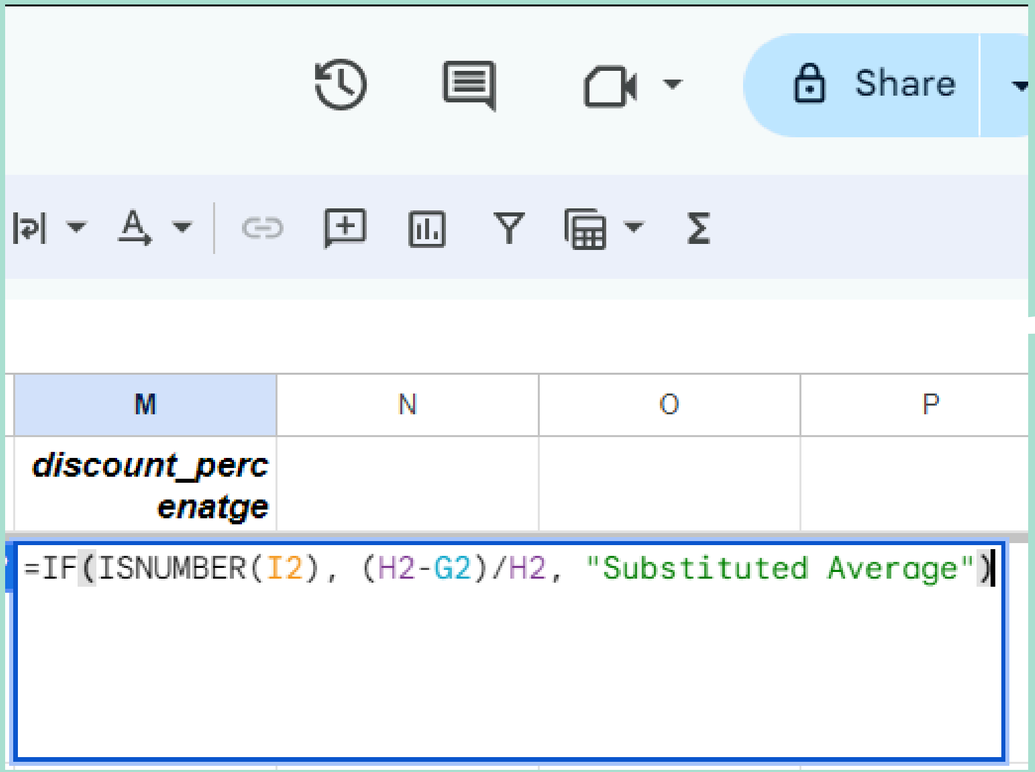
Once again we can utilise pivot tables to calculate averages.

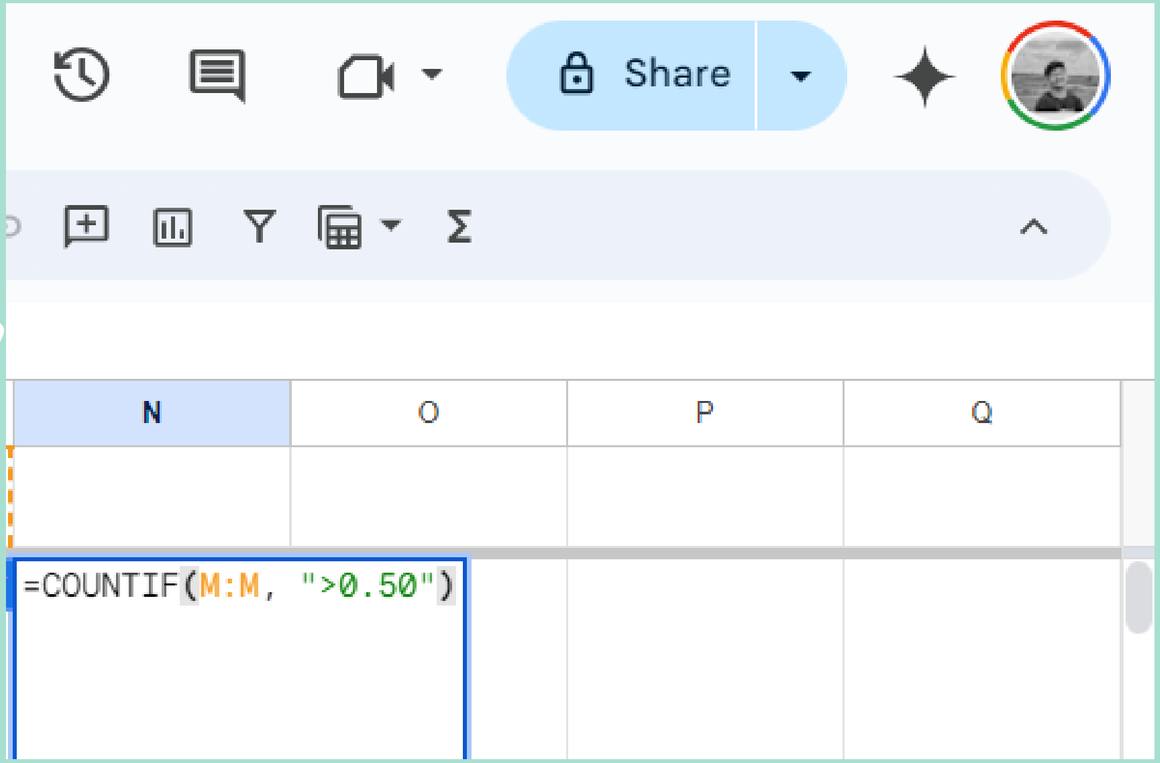
1. From rows dropdown select ratings.
2. From values drop down select OriginalPrice
3. Now from filter dropdown add ratings then select filter by condition and select greater than add 4 in the box below.



Number of products with a discount offer greater than 50% OFF.

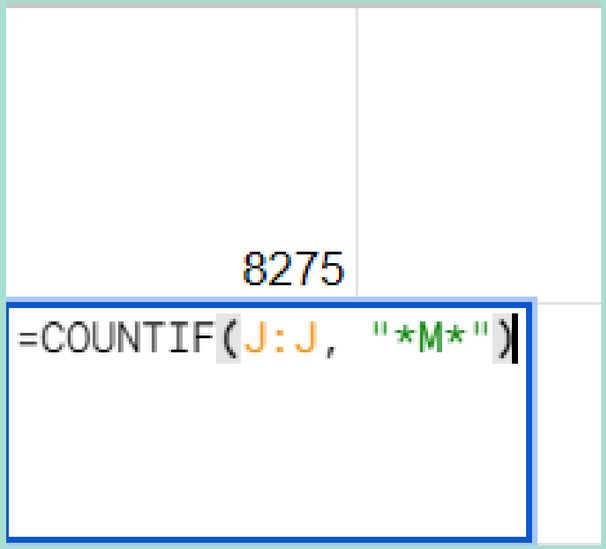
1. Calculate the discount percenatge
2. Use COUNTIF
3. Result: 8275





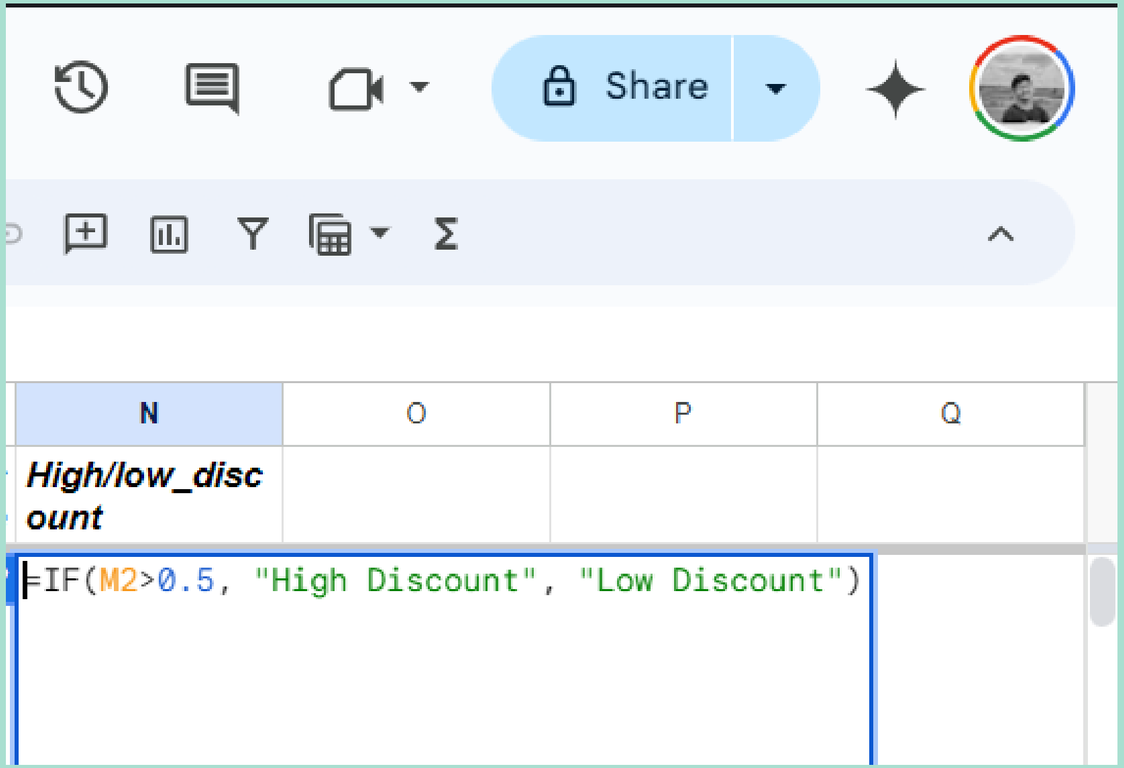
Number of products available in size "M."

Use COUNTIF with a wildcard pattern



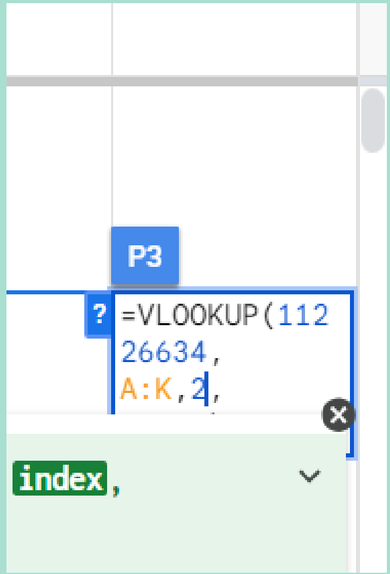
Labelling the products as “High Discount” or “Low Discount”

Use a simple IF statement

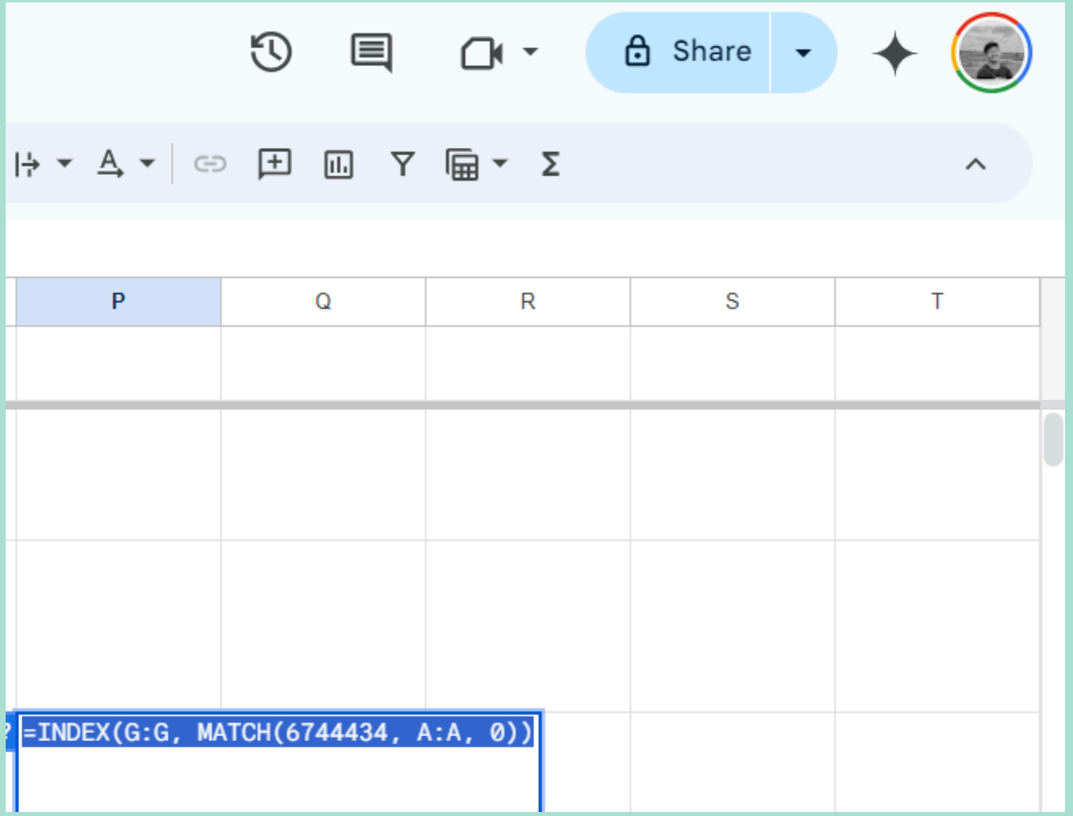


Use VLOOKUP/XLOOKUP for Data Retrival for a ProductID

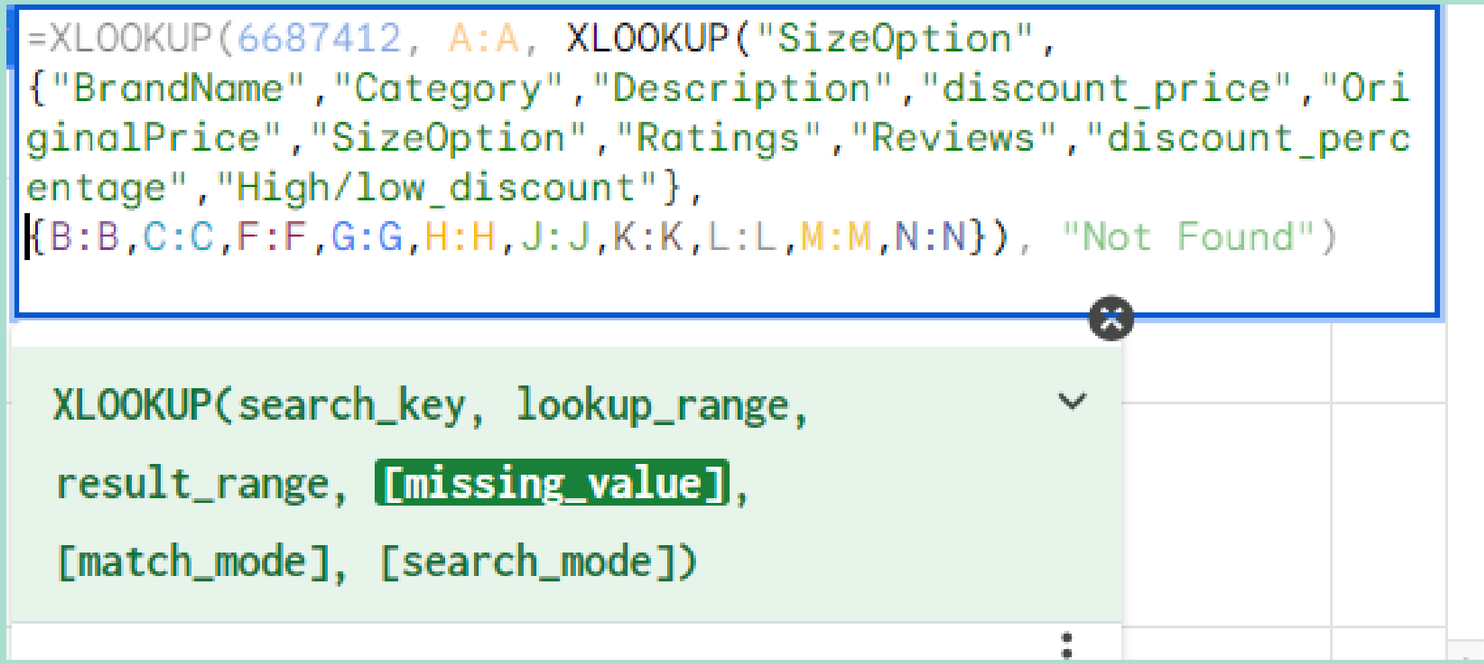
Use VLOOKUP to find the Brand, Price and Rating
Keep the range as A:I to only manipulate index.
Change the index to 2 for Brand, 8 for Price and 11 for Ratings.



Using MATCH and INDEX functions to find DiscountPrice of a product


Share
P
Q
R
S
T
=INDEX(G:G, MATCH(6744434, A:A, 0))

Using nested XLOOKUP to find all the columns of a product


=XLOOKUP(6687412, A:A, XLOOKUP("SizeOption",
{"BrandName","Category","Description","discount_price","Ori
ginalPrice","SizeOption","Ratings","Reviews","discount_perc
entage","High/low_discount"},
{B:B,C:C,F:F,G:G,H:H,J:J,K:K,L:L,M:M,N:N}), "Not Found")
XLOOKUP(search_key, lookup_range,
result_range, [missing_value],
[match_mode], [search_mode])

Thank you very much!