

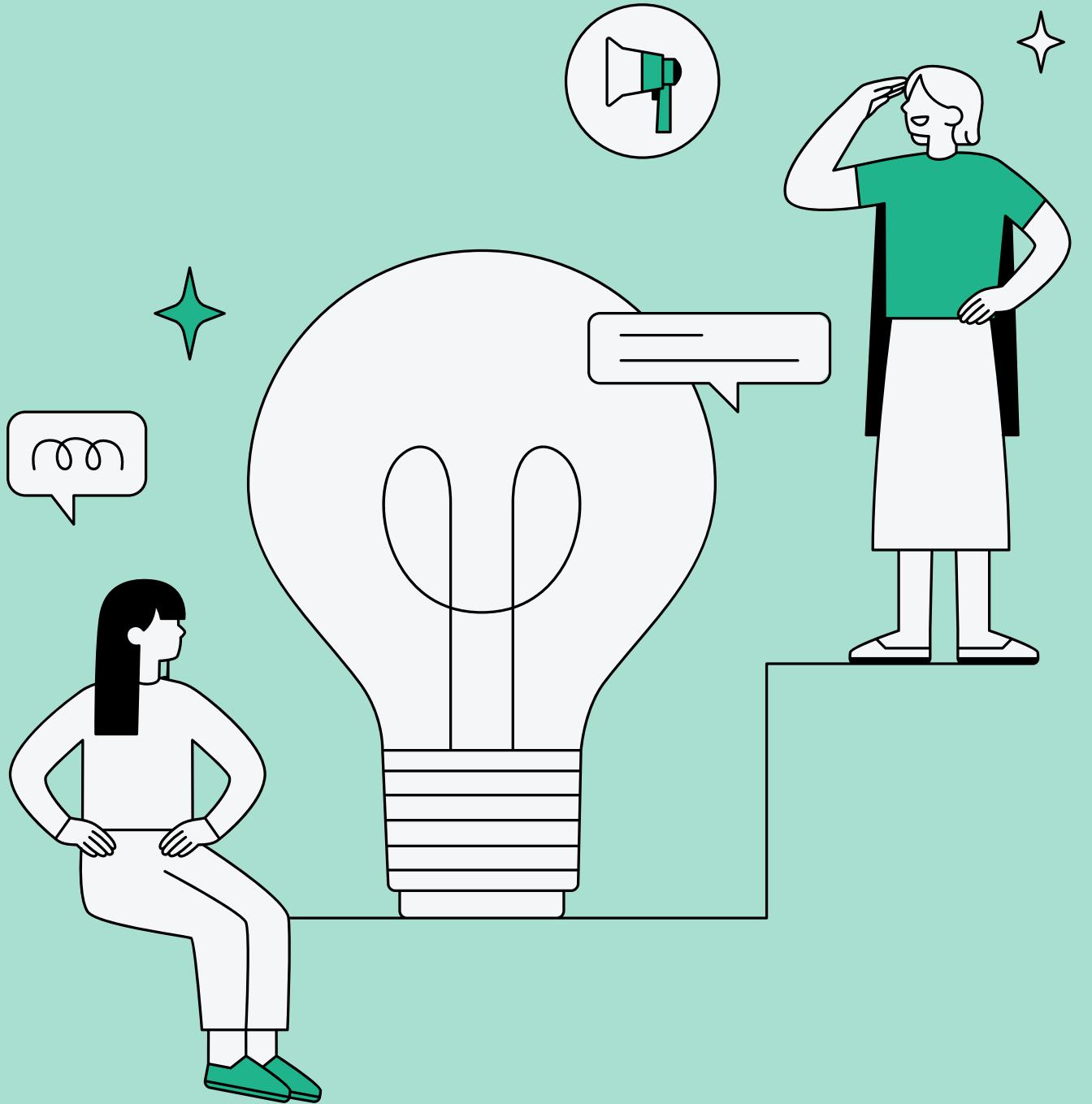
Presented by Azam
02-10-2024

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

mynta_apperial_dataset

File Edit View Insert Format Data Tools Extensions Help

A:A Product_id

A B C

1 Product_id BrandName Category

2 2296012 Roadster Bottom Wear men slim jeans ₹825.00 ₹1,499.00 ₹674.00 999 S, M, L, XL, XXL

3 13780156 LOCOMOTIVE Bottom Wear men track-pants ₹518.00 ₹1,149.00 ₹631.00 999 S, M, L, XL, XXL

4 11895958 Roadster Topwear men shirts ₹630.00 ₹1,399.00 ₹769.00 999 S, M, L, XL, XXL

5 4335679 Zivame Lingerie & Sleep Wear men shorts ₹390.00 ₹599.00 ₹209.00 999 XS, S, M, L, XL, XXL

+ mynta_dataset - Mynta Fashion Clothing



mynta_apperial_dataset

File Edit View Insert Format Data Tools Extensions Help

A2:A15001 2296012

A B C D E F G H I J K L M

1 Product_id BrandName Category Individual_cat category_by_Gender Description discounted_price (in Rs) OriginalPrice (in Rs) discount Reviews SizeOption

2 2296012 Roadster Bottom Wear jeans Men roadster men navy blue slim 28, 30, 32, 34, 36

3 13780156 LOCOMOTIVE Bottom Wear track-pants Men

4 11895958 Roadster Topwear shirts Men

5 4335679 Zivame Lingerie & Sleep Wear shapewear Women zivame women black saree shapewear zi3023core0nu de 894.00 1,295.00 401.00 999 S, M, L, XL, XXL

6 11690882 Roadster Western tshirts Women roadster women white solid v neck pure cotton t shirt mast harbour 390.00 599.00 209.00 999 XS, S, M, L, XL

Sum: 162,064,9

Remove duplicates

No duplicate rows were found.

15000 unique rows remain.

OK

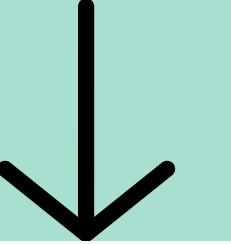
**Standardising
Discount Offer
Column**

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

1. We split the column on space delimiter

Screenshot of a Google Sheets interface showing a table titled "DiscountOffer". The table has columns: Product_id, BrandName, Category, Individual_cat, category_by_Gender, Description, DiscountPrice (in Rs), OriginalPrice (in Rs), DiscountOffer, SizeOption, Ratings, and Reviews. The "DiscountOffer" column contains values like "45% OFF", "55% OFF", and "31% OFF". A context menu is open over the "DiscountOffer" column, with the "Split text to columns" option highlighted.

Product_id	BrandName	Category	Individual_cat	category_by_Gender	Description	DiscountPrice (in Rs)	OriginalPrice (in Rs)	DiscountOffer	SizeOption	Ratings	Reviews
2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean look jeans	824	1499	45% OFF	28, 30, 32, 34, 36	3.9	999
13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit track pants	517	1149	55% OFF	S, M, L, XL	4	999
11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric printed sustainable casual shirt	629	1399	55% OFF	38, 40, 42, 44, 46, 48	4.3	999
4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0nude	893	1295	31% OFF	S, M, L, XL, XXL	4.2	999
11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton t shirt	599	999	35% OFF	mast harbour	4.5	999



Screenshot of a Google Sheets interface showing the same table after splitting the "DiscountOffer" column. The "DiscountOffer" column now contains the percentage values (e.g., "45% OFF", "55% OFF", "31% OFF"). A dropdown menu is open next to the "Separator" button, showing "Space" as the selected option.

Product_id	BrandName	Category	Individual_cat	category_by_Gender	Description	DiscountPrice (in Rs)	OriginalPrice (in Rs)	DiscountOffer	SizeOption	Ratings	Reviews
2296012	Roadster	Bottom Wear	jeans	Men	roadster men navy blue slim fit mid rise clean look jeans	824	1499	45% OFF	28, 30, 32, 34, 36	3.9	999
13780156	LOCOMOTIVE	Bottom Wear	track-pants	Men	locomotive men black white solid slim fit track pants	517	1149	55% OFF	28, 30, 32, 34, 36	4	999
11895958	Roadster	Topwear	shirts	Men	roadster men navy white black geometric printed sustainable casual shirt	629	1399	55% OFF	28, 30, 32, 34, 36	4.3	999
4335679	Zivame	Lingerie & Sleep Wear	shapewear	Women	zivame women black saree shapewear zi3023core0nude	893	1295	31% OFF	28, 30, 32, 34, 36	4.2	999
11690882	Roadster	Western	tshirts	Women	roadster women white solid v neck pure cotton t shirt	599	999	35% OFF	28, 30, 32, 34, 36	4.5	999

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_cat	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(VALUE(I2)>1, VALUE(I2)/100, VALUE(I2))				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 mast harbour	599	35% OFF	0.35						999

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

Mynta Fashion Clothing

File Edit View Insert Format Data Tools Extensions Help

Share

L2 | `=IF(ISNUMBER(J2), J2, H2*K2)`

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Product_id	BrandName	Category	Individual_cat egory	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	674.55	=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									

+ Mynta Fashion Clothing

ENG IN 23:27 02-10-2024

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

The screenshot shows a Google Sheets document titled "Myntra Fashion Clothing". The sheet contains a table with columns: F (Description), G (DiscountPrice (in Rs)), H (OriginalPrice (in Rs)), I (discount_offer), and J (Reviews). The table lists various products with their descriptions, discounted prices, original prices, discount offers, and review counts. The "discount_offer" column for the first product is highlighted with a blue border. The "Insert" menu is open, and the "Pivot table" option is selected. The status bar at the bottom right shows "Sum: 162,118,312,93..." and the date "02-10-2024".

y_by_	F	G	H	I	J	K	L	M	N	O
	Description	DiscountPrice (in Rs)	OriginalPrice (in Rs)	discount_offer	Reviews					
1	Product_id	Brand								
2	2296012	Roadster men navy blue slim fit mid rise clean look jeans	₹824.00	₹1,499.00	₹674.55	999				
3	13780156	Locomotive men black white solid slim fit track pants	₹517.00	₹1,149.00	₹631.95	999				
4	11895958	Roadster men navy white black geometric printed sustainable casual shirt	₹629.00	₹1,399.00	₹769.45	999				
5	4335679	Zivame women black saree shapewear zi3023core0nude	₹893.00	₹1,295.00	₹401.45	999				
6	11690882	Roadster women white solid v neck pure cotton t shirt mast harbour		₹599.00	₹209.65	999				

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

The image consists of three sequential screenshots of the Google Sheets Pivot Table Editor, illustrating the steps to create a pivot table:

- Step 1:** The first screenshot shows the "Rows" section of the editor. A white arrow points from the "Rows" section to the "Add" button. The "Category" field is selected in the dropdown menu.
- Step 2:** The second screenshot shows the "Values" section of the editor. A white arrow points from the "Values" section to the "Add" button. The "discount_price" field is selected in the dropdown menu.
- Step 3:** The third screenshot shows the "Values" section again, but now the "Summarise by" dropdown menu is open. A white arrow points to the "AVERAGE" option, which is highlighted.

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

The screenshot shows a Google Sheets interface with a toolbar at the top. Below the toolbar is a menu bar with icons for file, edit, insert, format, data, and more. The main area displays a table with columns labeled K, L, M, N, and O. In cell K2, there is a value of ₹824.45 and a formula: =IF(AND(ISBLANK(G2), ISBLANK(I2)), VLOOKUP(C2, 'Pivot Table'!\$A\$1:\$B\$10, 2, FALSE), G2). The formula is highlighted with a blue border.

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

A screenshot of a Google Sheets spreadsheet. The top menu bar includes icons for clock, comment, refresh, share (with a user profile picture), and other options. Below the menu is a toolbar with icons for bold, italic, underline, font color, font size, and more. The main area shows a table with columns labeled K, L, M, N, and O. Cell K2 contains the value ₹824.45. Cell K3 contains the formula =IF(AND(ISBLANK(G2), ISBLANK(I2)), VLOOKUP(C2, 'Pivot Table'!\$A\$1:\$B\$10, 2, FALSE), G2). The formula is highlighted with a blue border.

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.

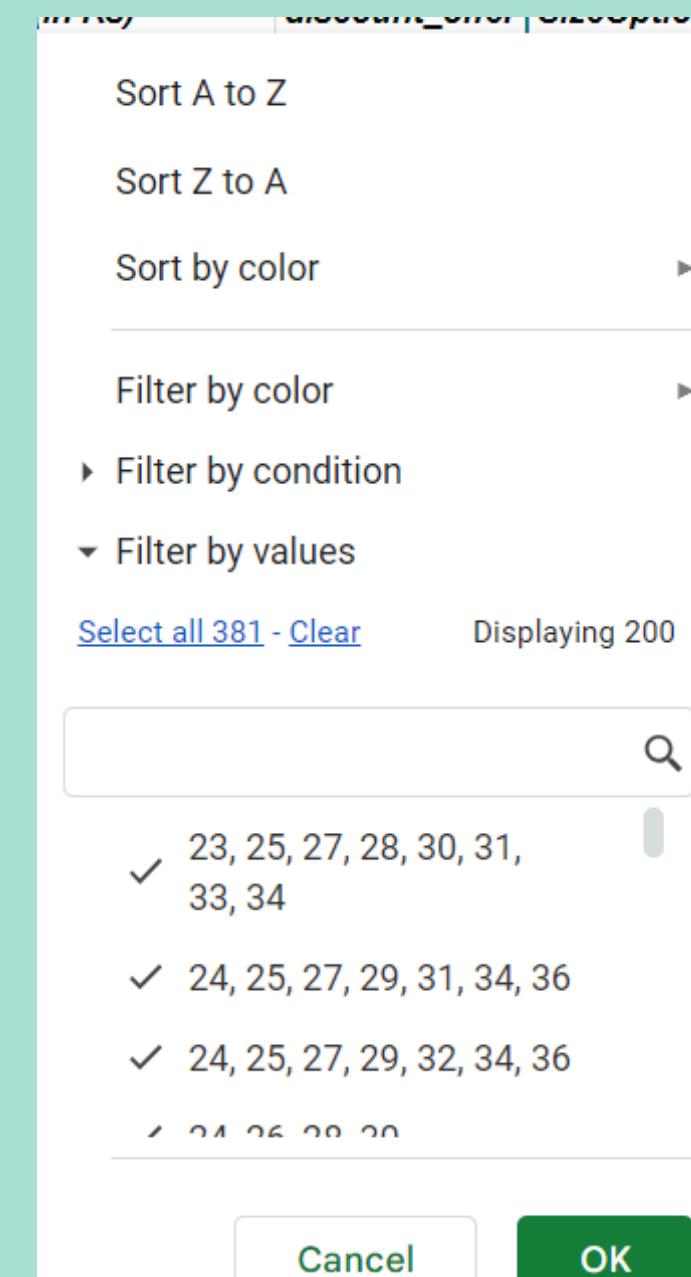
1. Select the complete column
2. Click the filter icon and check for
(Blanks)

The image shows three screenshots of Microsoft Excel illustrating the process of identifying blank values in a column:

- Initial View:** A screenshot of an Excel spreadsheet with columns J, K, and L. The header row contains "SizeOption", "Ratings", and "Reviews". The data rows show various size options like "28, 30, 32, 34, 3" and "S, M, L, XL".
- Filter Icon Click:** A screenshot showing the Excel ribbon with the "Filter" icon (a funnel symbol) highlighted. A large white arrow points from the initial view to this step.
- Filtered Results:** A screenshot showing the same spreadsheet after applying the filter. The "SizeOption" column now highlights rows where the value is blank or null. The first two rows ("28, 30, 32, 34, 3" and "S, M, L, XL") are no longer visible, while the others ("38, 40, 42, 44, 4", "S, M, L, XL, XXL", and "XS, S, M, L, XL") remain.

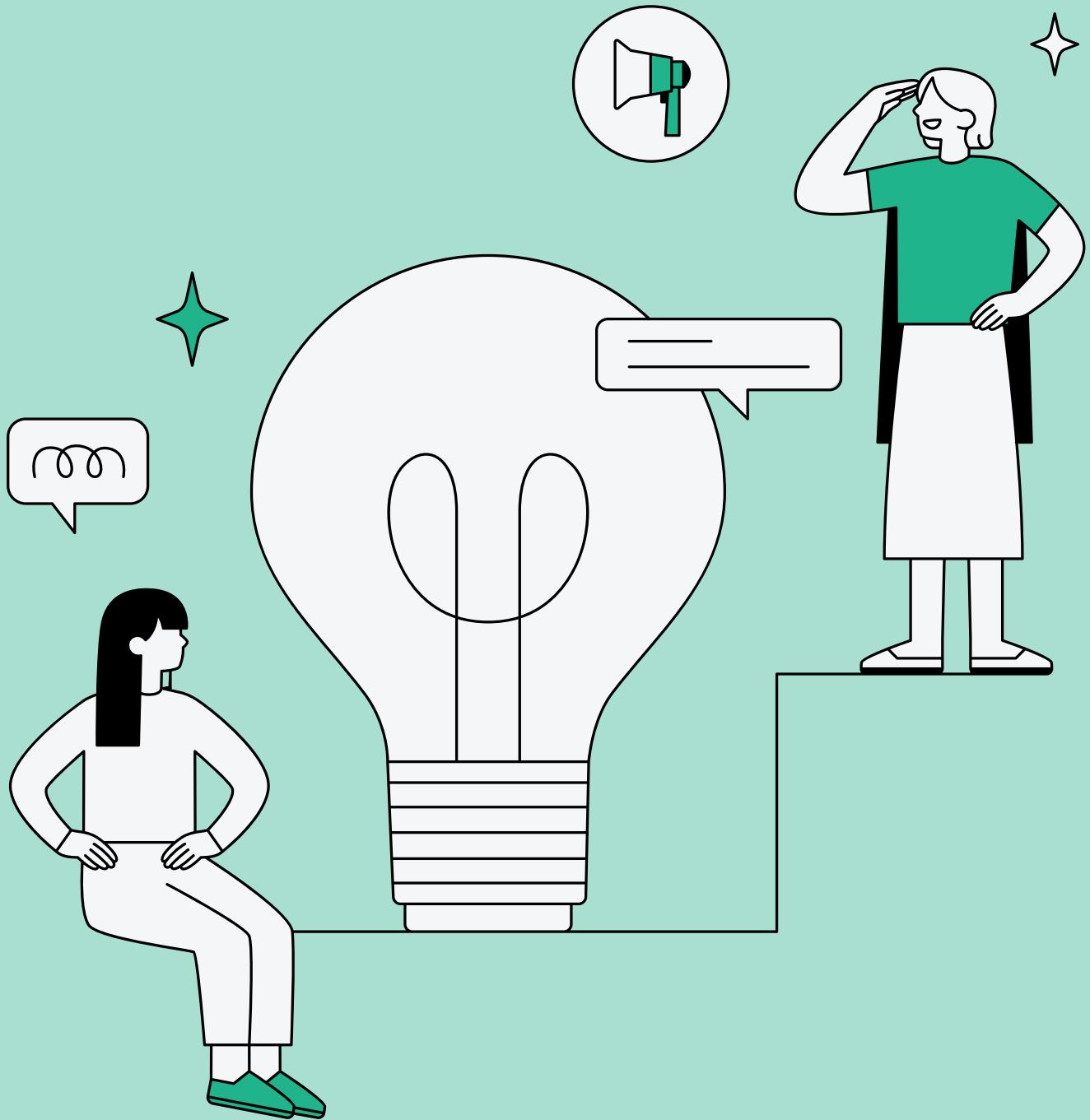
SizeOption	Ratings	Reviews
28, 30, 32, 34, 3	3.9	999
S, M, L, XL	4	999
38, 40, 42, 44, 4	4.3	999
S, M, L, XL, XXL	4.2	999
XS, S, M, L, XL	4.2	999

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.

The screenshot shows a Google Sheets document titled "Myntra Fashion Clothing". The main sheet displays a pivot table with the following data:

Ratings	AVERAGE of OriginalPrice (in Rs)
4.1	₹1,739.25
4.2	₹1,739.92
4.3	₹1,673.45
4.4	₹1,621.43
4.5	₹1,430.18
4.6	₹1,511.38
4.7	₹1,244.83
4.8	₹2,798.00
Grand Total	₹1,675.41

The "Pivot Table editor" sidebar is open on the right, showing the configuration details:

- Ratings**: Order Ascend..., Sort by Ratings, Show totals checked.
- Columns**: Add button.
- Values**: OriginalPrice (in Rs) selected, Summarize by AVERA..., Show as Default.
- Filters**: Ratings selected, Status Value is greater than 4.
- Search**: Product_id, BrandName, Category, Individual_category, category_by_Gender, Description, discount_price, OriginalPrice (in Rs), discount_offer, SizeOption, Ratings, Reviews.

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

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

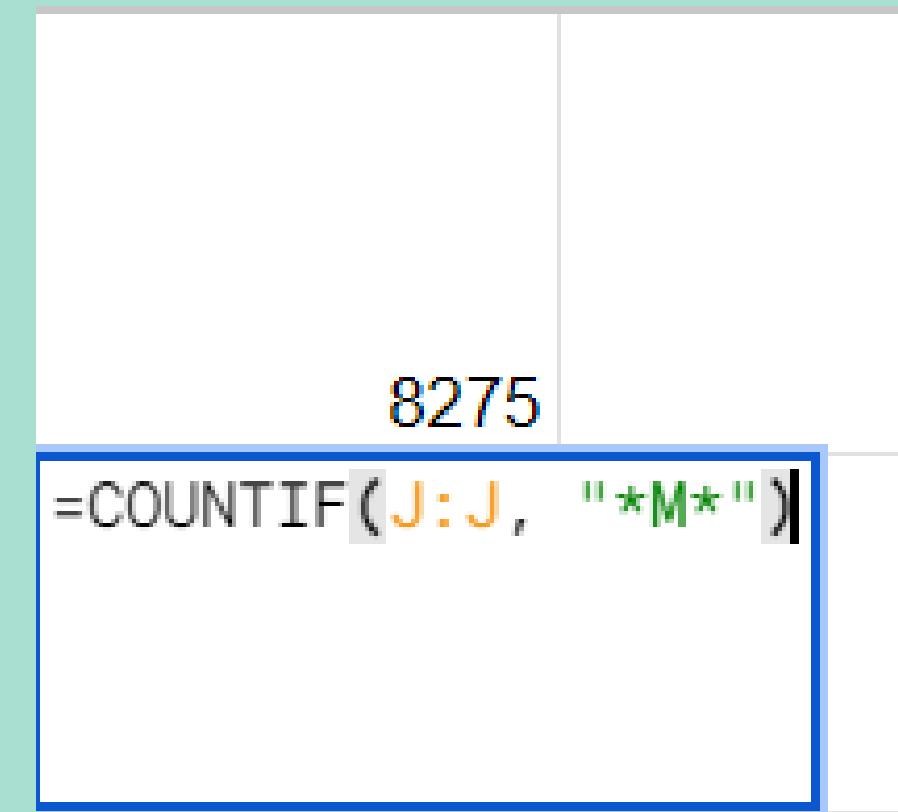


M	N	O	P
discount_percent			
=IF(ISNUMBER(I2), (H2-G2)/H2, "Substituted Average")			

N	O	P	Q
=COUNTIF(M:M, ">0.50")			

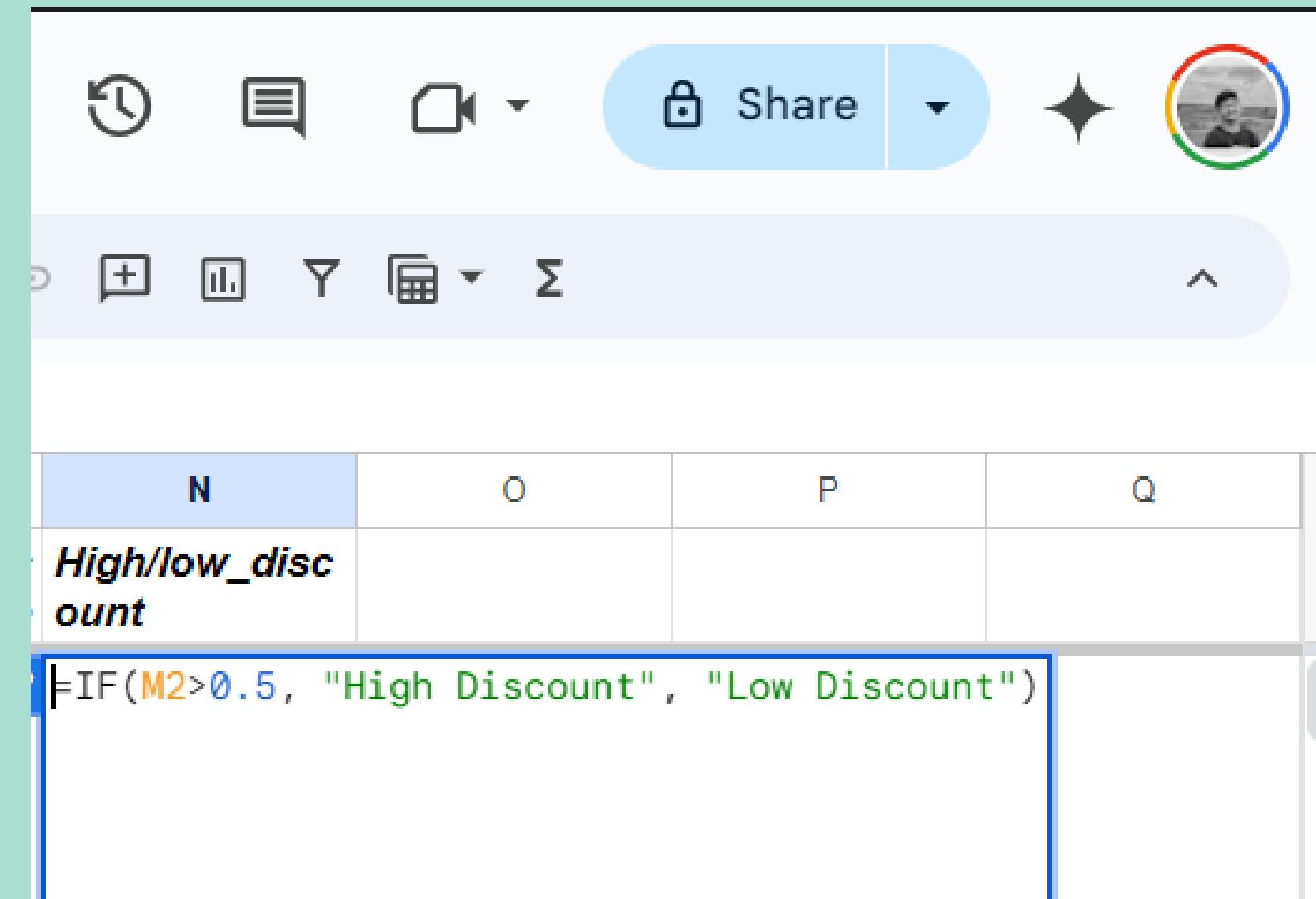
Number of products
available in size "M."

Use COUNTIF with a wildcard pattern



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

Use a simple IF statement



The screenshot shows a portion of a Google Sheets interface. At the top, there's a toolbar with icons for clock, document, video, share (highlighted in blue), and user profile. Below the toolbar is a menu bar with icons for file, insert, filter, sort, calculate, and formulas. The main area displays a table with four columns labeled N, O, P, and Q. The first row contains the header 'High/low_discount'. The second row contains the formula '=IF(M2>0.5, "High Discount", "Low Discount")' in cell N2. The formula uses the M2 cell reference, which is highlighted in orange.

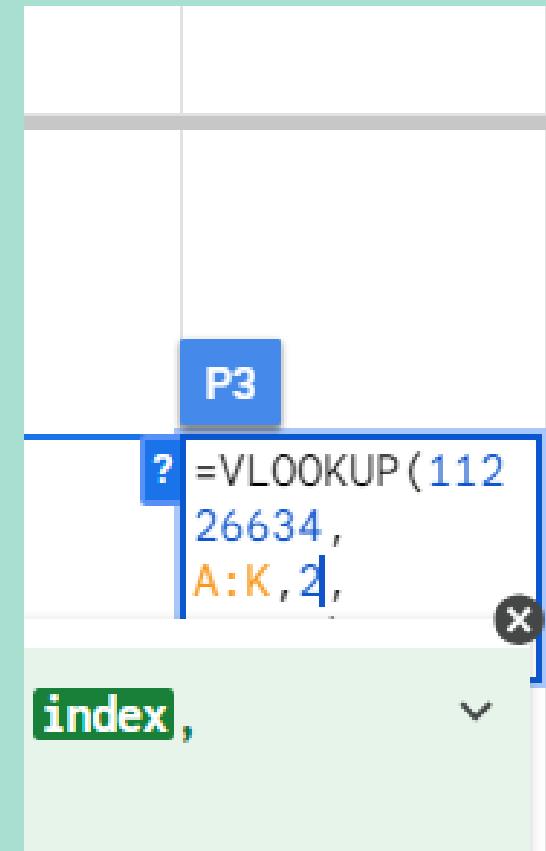
N	O	P	Q
High/low_discount			
=IF(M2>0.5, "High Discount", "Low Discount")			

Use
VLOOKUP/XLOOKUP
for Data Retrieval for a
ProductID

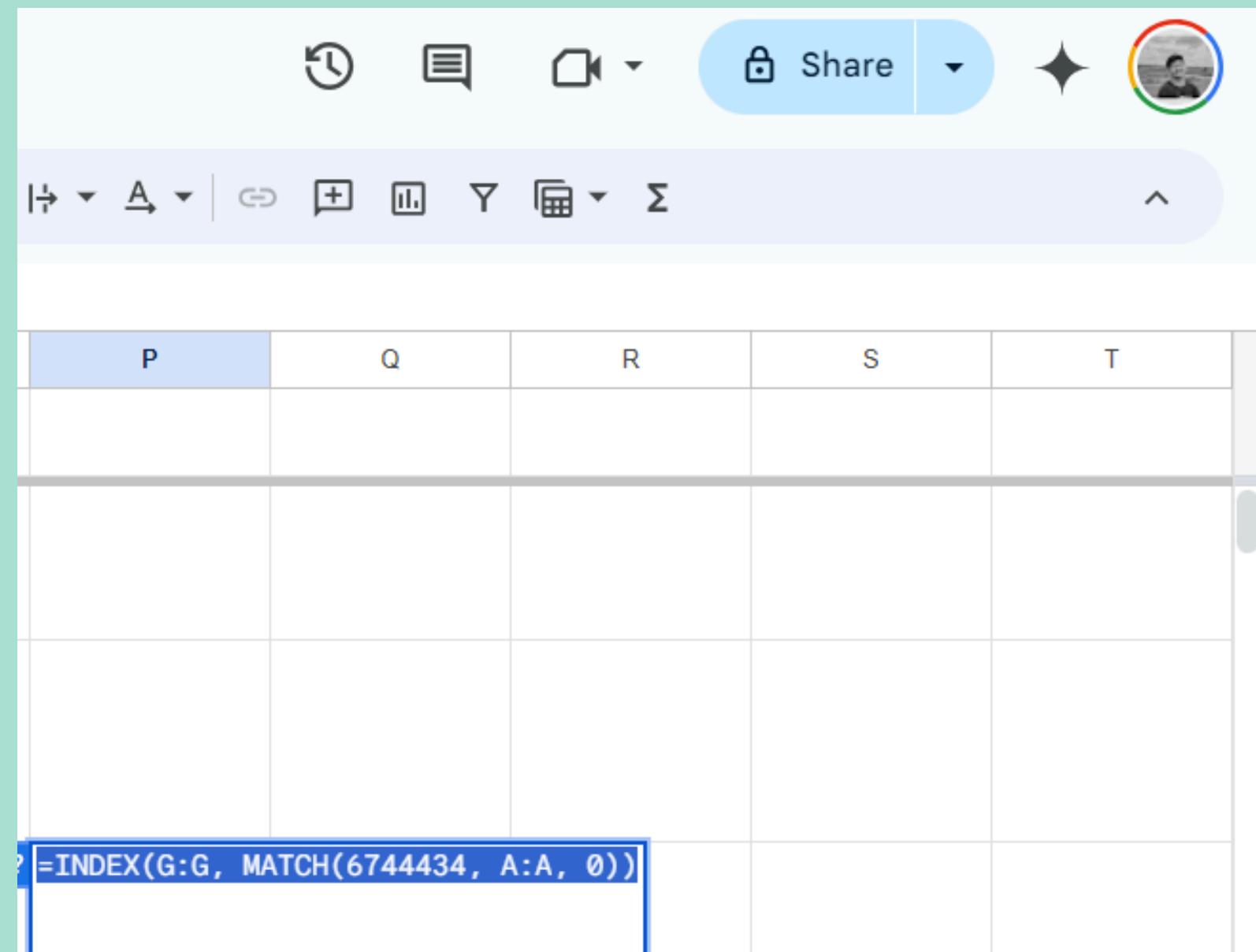
Use VLOOKUP to find the Brand, Price and Rating

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



Using nested
XLOOKUP to find all the
columns of a product

```
=XLOOKUP(6687412, A:A, XLOOKUP("SizeOption",  
{"BrandName", "Category", "Description", "discount_price", "OriginalPrice", "SizeOption", "Ratings", "Reviews", "discount_percentage", "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!

