SUPERSTORE

François SOK

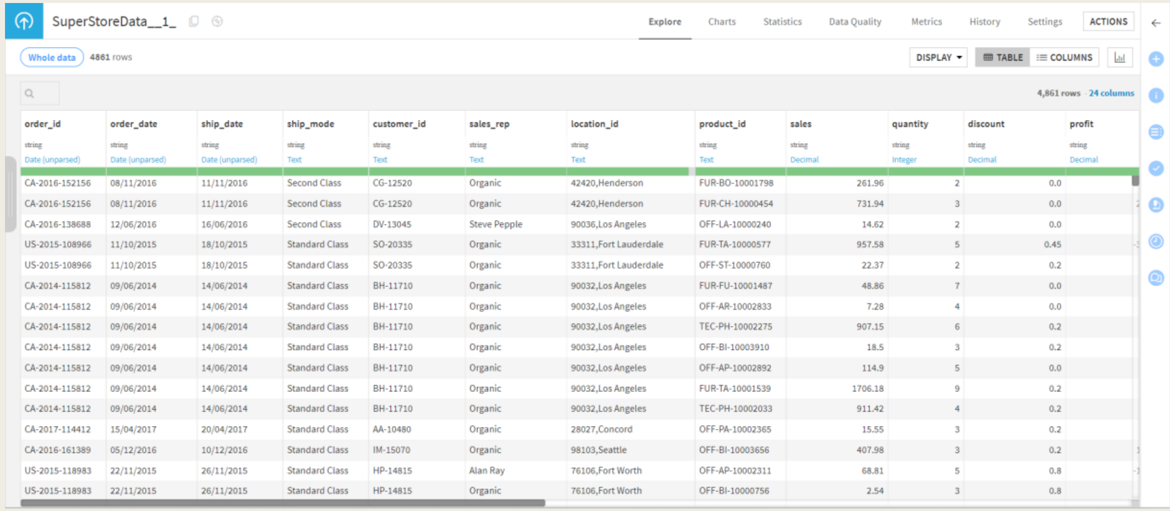
1. Collecte et préparation des données

Légende:

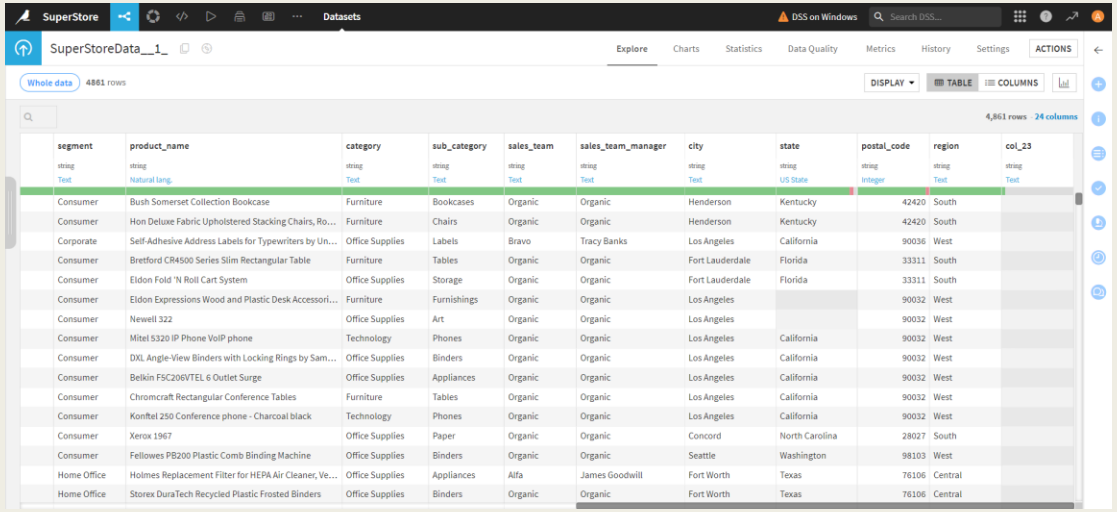
- L’exposition du problème rencontré

→ En vert, les solutions proposées

A. Détection des anomalies



order_id	order_date	ship_date	ship_mode	customer_id	sales_rep	location_id	product_id	sales	quantity	discount	profit
string	string	string	string	string	string	string	string	string	string	string	string
Date (unparsed)	Date (unparsed)	Date (unparsed)	Text	Text	Text	Text	Text	Decimal	Integer	Decimal	Decimal
CA-2016-152156	08/11/2016	11/11/2016	Second Class	CG-12520	Organic	42420,Henderson	FUR-BO-10001798	261.96	2	0.0	
CA-2016-152156	08/11/2016	11/11/2016	Second Class	CG-12520	Organic	42420,Henderson	FUR-CH-10000454	731.94	3	0.0	
CA-2016-138688	12/06/2016	16/06/2016	Second Class	DV-13045	Steve Pepple	90036,Los Angeles	OFF-LA-10000240	14.62	2	0.0	
US-2015-108966	11/10/2015	18/10/2015	Standard Class	SO-20335	Organic	33311,Fort Lauderdale	FUR-TA-10000577	957.58	5	0.45	
US-2015-108966	11/10/2015	18/10/2015	Standard Class	SO-20335	Organic	33311,Fort Lauderdale	OFF-ST-10000760	22.37	2	0.2	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	FUR-FU-10001487	48.86	7	0.0	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	OFF-AR-10002833	7.28	4	0.0	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	TEC-PH-10002275	907.15	6	0.2	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	OFF-BI-10003910	18.5	3	0.2	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	OFF-AP-10002892	114.9	5	0.0	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	FUR-TA-10001539	1706.18	9	0.2	
CA-2014-115812	09/06/2014	14/06/2014	Standard Class	BH-11710	Organic	90032,Los Angeles	TEC-PH-10002033	911.42	4	0.2	
CA-2017-114412	15/04/2017	20/04/2017	Standard Class	AA-10480	Organic	28027,Concord	OFF-PA-10002365	15.55	3	0.2	
CA-2016-161389	05/12/2016	10/12/2016	Standard Class	IM-15070	Organic	98103,Seattle	OFF-BI-10003656	407.98	3	0.2	
US-2015-118983	22/11/2015	26/11/2015	Standard Class	HP-14815	Alan Ray	76106,Fort Worth	OFF-AP-10002311	68.81	5	0.8	
US-2015-118983	22/11/2015	26/11/2015	Standard Class	HP-14815	Organic	76106,Fort Worth	OFF-BI-10000756	2.54	3	0.8	



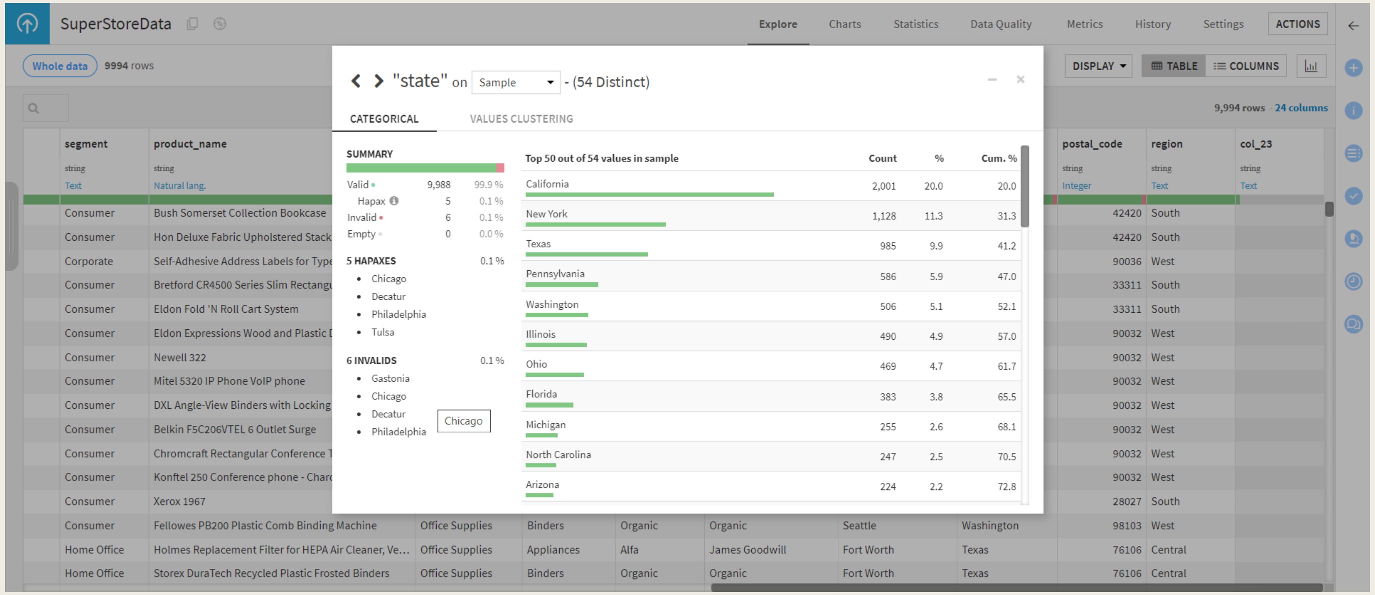
segment	product_name	category	sub_category	sales_team	sales_team_manager	city	state	postal_code	region	col_23
string	string	string	string	string	string	string	string	string	string	string
Text	Natural lang.	Text	Text	Text	Text	Text	US State	Integer	Text	Text
Consumer	Bush Somerset Collection Bookcase	Furniture	Bookcases	Organic	Organic	Henderson	Kentucky	42420	South	
Consumer	Hon Deluxe Fabric Upholstered Stacking Chairs, Ro...	Furniture	Chairs	Organic	Organic	Henderson	Kentucky	42420	South	
Corporate	Self-Adhesive Address Labels for Typewriters by Un...	Office Supplies	Labels	Bravo	Tracy Banks	Los Angeles	California	90036	West	
Consumer	Bretford CR4500 Series Slim Rectangular Table	Furniture	Tables	Organic	Organic	Fort Lauderdale	Florida	33311	South	
Consumer	Eldon Fold 'N Roll Cart System	Office Supplies	Storage	Organic	Organic	Fort Lauderdale	Florida	33311	South	
Consumer	Eldon Expressions Wood and Plastic Desk Accessori...	Furniture	Furnishings	Organic	Organic	Los Angeles		90032	West	
Consumer	Newell 322	Office Supplies	Art	Organic	Organic	Los Angeles		90032	West	
Consumer	Mitel 5320 IP Phone VoIP phone	Technology	Phones	Organic	Organic	Los Angeles	California	90032	West	
Consumer	DXL Angle-View Binders with Locking Rings by Sam...	Office Supplies	Binders	Organic	Organic	Los Angeles	California	90032	West	
Consumer	Belkin F5C206VTEL 6 Outlet Surge	Office Supplies	Appliances	Organic	Organic	Los Angeles	California	90032	West	
Consumer	Chromcraft Rectangular Conference Tables	Furniture	Tables	Organic	Organic	Los Angeles	California	90032	West	
Consumer	Konftel 250 Conference phone - Charcoal black	Technology	Phones	Organic	Organic	Los Angeles	California	90032	West	
Consumer	Xerox 1967	Office Supplies	Paper	Organic	Organic	Concord	North Carolina	28027	South	
Consumer	Fellowes PB200 Plastic Comb Binding Machine	Office Supplies	Binders	Organic	Organic	Seattle	Washington	98103	West	
Home Office	Holmes Replacement Filter for HEPA Air Cleaner, Ve...	Office Supplies	Appliances	Alfa	James Goodwill	Fort Worth	Texas	76106	Central	
Home Office	Storex DuraTech Recycled Plastic Frosted Binders	Office Supplies	Binders	Organic	Organic	Fort Worth	Texas	76106	Central	

Anomalies détectées dans les colonnes suivantes:

- location_id,
- state,
- postal_code,
- region et
- col_23.

i. Valeurs invalides

- Colonnes concernées: product_name, category, sub_category, sales_team, sales_team_manager, city, state, postal_code, region et col_23
  - *Concerne uniquement les lignes dont la « Category » est « 16GB ».*



"state" on Sample - (54 Distinct)

CATEGORICAL   VALUES CLUSTERING

SUMMARY

Valid	9,988	99.9 %
Hapax	5	0.1 %
Invalid	6	0.1 %
Empty	0	0.0 %

5 HAPAXES 0.1 %

- Chicago
- Decatur
- Philadelphia
- Tulsa

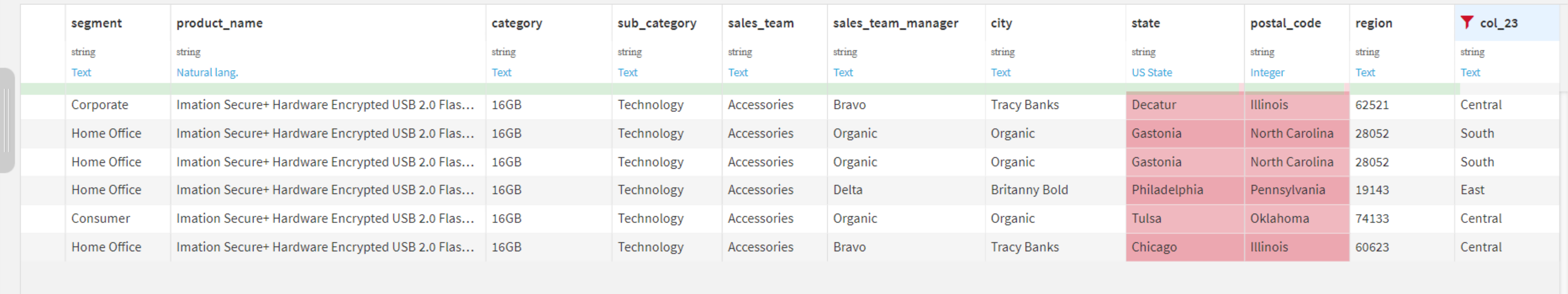
6 INVALIDS 0.1 %

- Gastonia
- Chicago
- Decatur
- Philadelphia

Chicago

Top 50 out of 54 values in sample	Count	%	Cum. %
California	2,001	20.0	20.0
New York	1,128	11.3	31.3
Texas	985	9.9	41.2
Pennsylvania	586	5.9	47.0
Washington	506	5.1	52.1
Illinois	490	4.9	57.0
Ohio	469	4.7	61.7
Florida	383	3.8	65.5
Michigan	255	2.6	68.1
North Carolina	247	2.5	70.5
Arizona	224	2.2	72.8

segment string Text	product_name string Natural lang.							postal_code string Integer	region string Text	col_23 string Text
Consumer	Bush Somerset Collection Bookcase							42420	South	
Consumer	Hon Deluxe Fabric Upholstered Stack							42420	South	
Corporate	Self-Adhesive Address Labels for Type							90036	West	
Consumer	Bretford CR4500 Series Slim Rectangu							33311	South	
Consumer	Eldon Fold 'N Roll Cart System							33311	South	
Consumer	Eldon Expressions Wood and Plastic I							90032	West	
Consumer	Newell 322							90032	West	
Consumer	Mitel 5320 IP Phone VoIP phone							90032	West	
Consumer	DXL Angle-View Binders with Locking							90032	West	
Consumer	Belkin F5C206VTEL 6 Outlet Surge							90032	West	
Consumer	Chromcraft Rectangular Conference T							90032	West	
Consumer	Konftel 250 Conference phone - Charc							90032	West	
Consumer	Xerox 1967							28027	South	
Consumer	Fellowes PB200 Plastic Comb Binding Machine	Office Supplies	Binders	Organic	Organic	Seattle	Washington	98103	West	
Home Office	Holmes Replacement Filter for HEPA Air Cleaner, Ve…	Office Supplies	Appliances	Alfa	James Goodwill	Fort Worth	Texas	76106	Central	
Home Office	Storex DuraTech Recycled Plastic Frosted Binders	Office Supplies	Binders	Organic	Organic	Fort Worth	Texas	76106	Central	

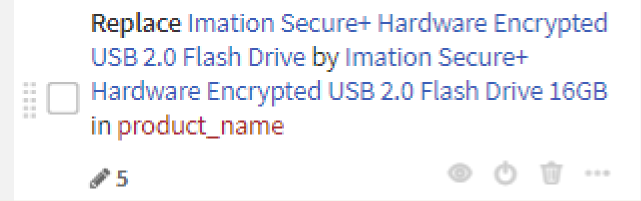


segment string Text	product_name string Natural lang.	category string Text	sub_category string Text	sales_team string Text	sales_team_manager string Text	city string Text	state string US State	postal_code string Integer	region string Text	col_23 string Text
Corporate	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Bravo	Tracy Banks	Decatur	Illinois	62521	Central
Home Office	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Organic	Organic	Gastonia	North Carolina	28052	South
Home Office	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Organic	Organic	Gastonia	North Carolina	28052	South
Home Office	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Delta	Britanny Bold	Philadelphia	Pennsylvania	19143	East
Consumer	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Organic	Organic	Tulsa	Oklahoma	74133	Central
Home Office	Imation Secure+ Hardware Encrypted USB 2.0 Flas…	16GB	Technology	Accessories	Bravo	Tracy Banks	Chicago	Illinois	60623	Central

i. Valeurs invalides - suite

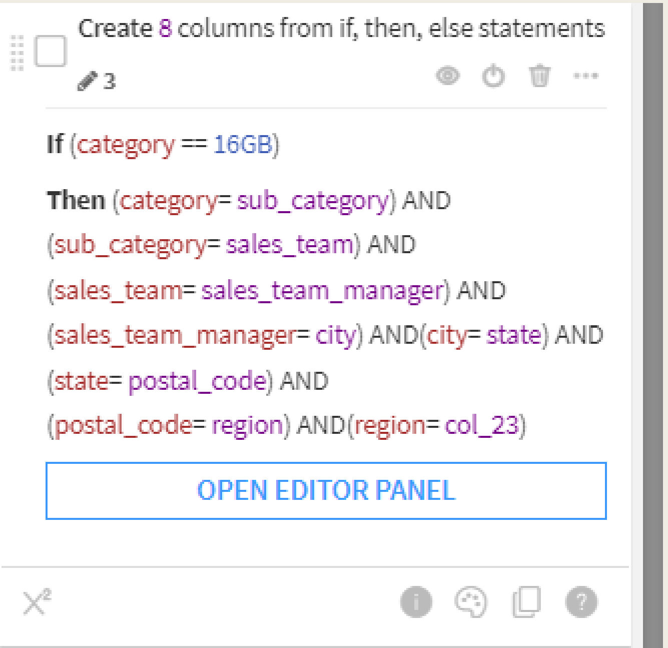
- Constat: il y a un décalage des colonnes.

→ product_name: **remplacer** « Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive » par « Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive 16GB ». En gros, on rajoute « 16GB » à la fin du texte.



Replace Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive by Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive 16GB in product_name

→ Pour corriger ce décalage, on pourrait faire un « shift ». Seulement, la fonction n'existe pas dans Dataiku donc on utilise la méthode suivante:



Create 8 columns from if, then, else statements

If (category == 16GB)

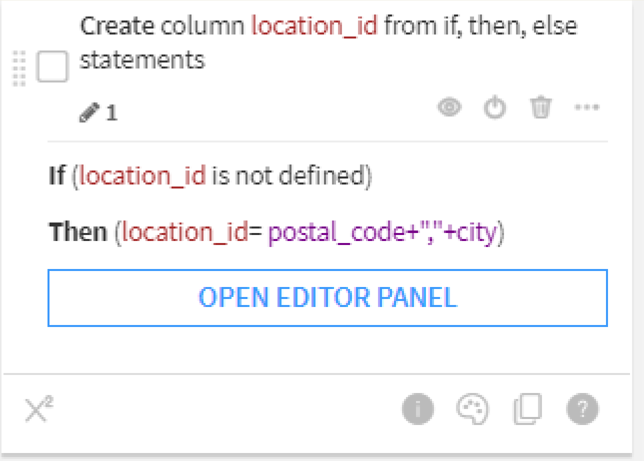
Then (category= sub_category) AND (sub_category= sales_team) AND (sales_team= sales_team_manager) AND (sales_team_manager= city) AND(city= state) AND (state= postal_code) AND (postal_code= region) AND(region= col_23)

ii. Valeurs manquantes

- Colonnes concernées: location_id et state

→ location_id: corrigés via la formule de la capture d’écran ci-dessous. Etape facultative (voir « iii. Doublons »

→ state: : corrigés en remplaçant par les valeurs trouvées dans « location_id » et en faisant le lien avec « postal_code »

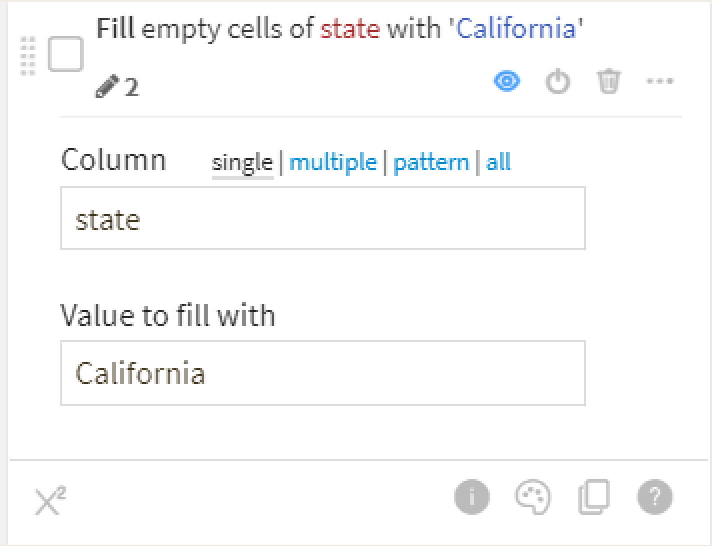


□ Create column location_id from if, then, else statements

1

If (location_id is not defined)

Then (location_id= postal_code+","+city)



□ Fill empty cells of state with 'California'

2

Column: state

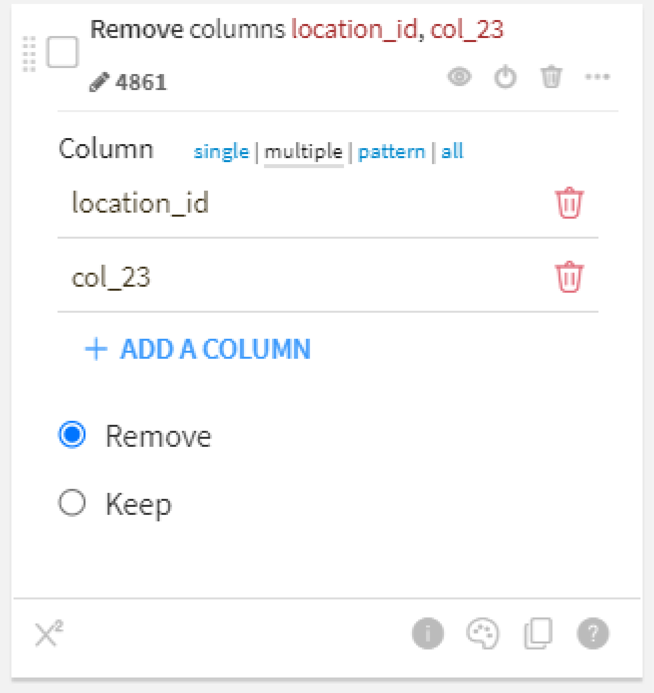
Value to fill with: California

iii. Doublons

- Colonnes concernées: col_23 et location_id

→ Supprimer « col_23 » car a servi uniquement pour remplacer les chiffres dans « region »

→ Supprimer « location_id » car doublon avec les colonnes: city, state, postal_code et region



Remove columns location_id, col_23

4861

Column single | multiple | pattern | all

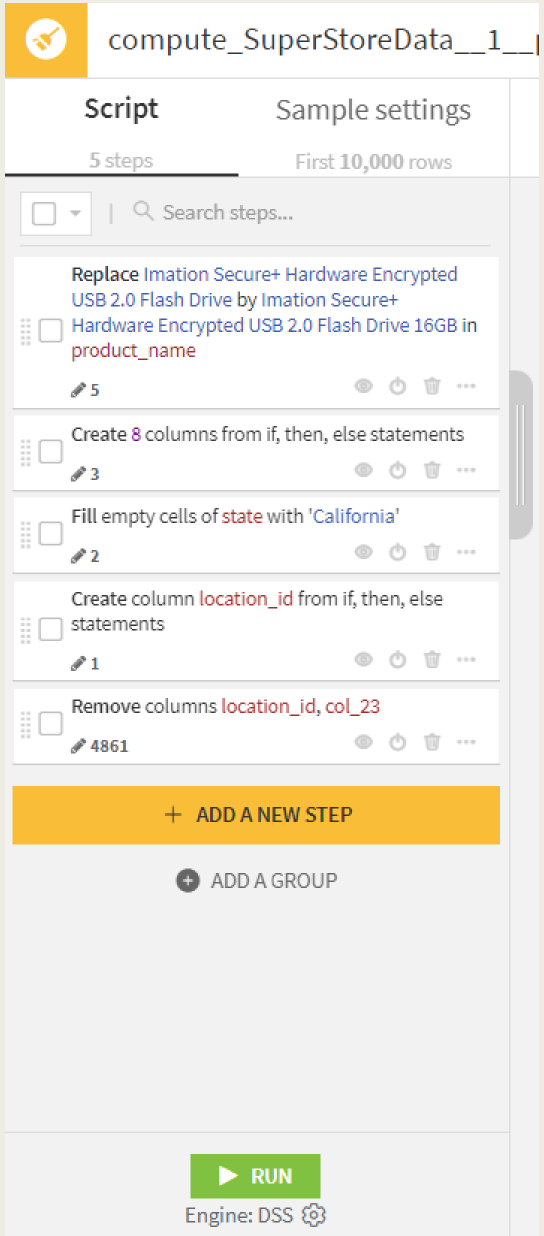
location_id

col_23

● Remove

○ Keep

Résumé des étapes dans Dataiku



compute_SuperStoreData__1__

Replace Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive by Imation Secure+ Hardware Encrypted USB 2.0 Flash Drive 16GB in product_name

5

Create 8 columns from if, then, else statements

3

Fill empty cells of state with 'California'

2

Create column location_id from if, then, else statements

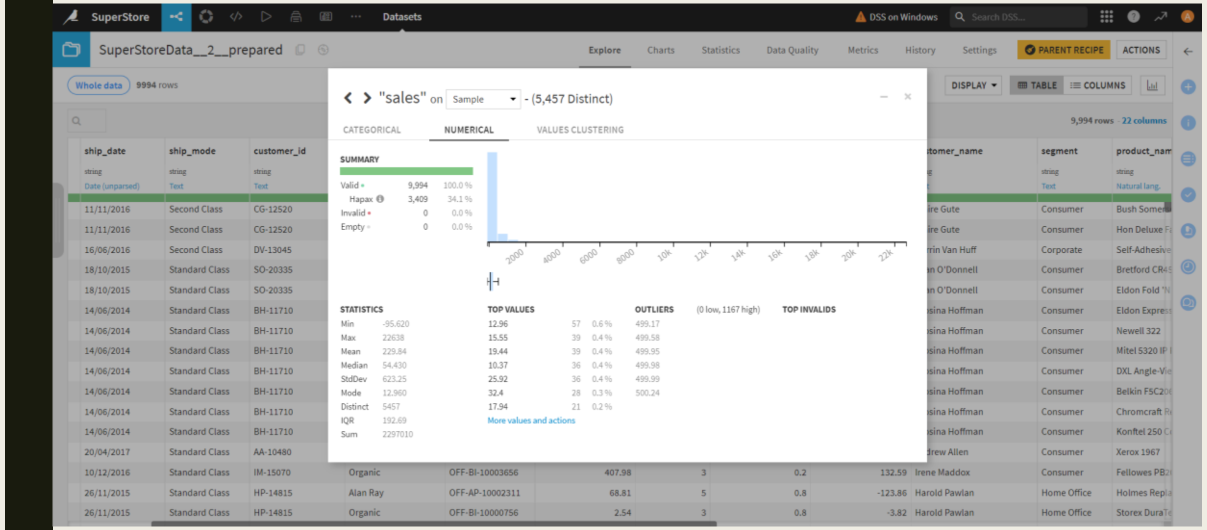
1

Remove columns location_id, col_23

4861

iv. Valeurs aberrantes

- Colonnes concernées: sales et profit
- sales et profit: les valeurs min. et max. sont bizarres, il faut investiguer.



SuperStoreData__2__prepared — Whole data 9994 rows

"sales" on Sample - (5,457 Distinct)

CATEGORICAL | NUMERICAL | VALUES CLUSTERING

SUMMARY

Valid	9,994	100.0 %
Hapax	3,409	34.1 %
Invalid	0	0.0 %
Empty	0	0.0 %

STATISTICS

Min	-95.620
Max	22638
Mean	229.84
Median	54.430
StdDev	623.25
Mode	12.960
Distinct	5457
IQR	192.69
Sum	2297010

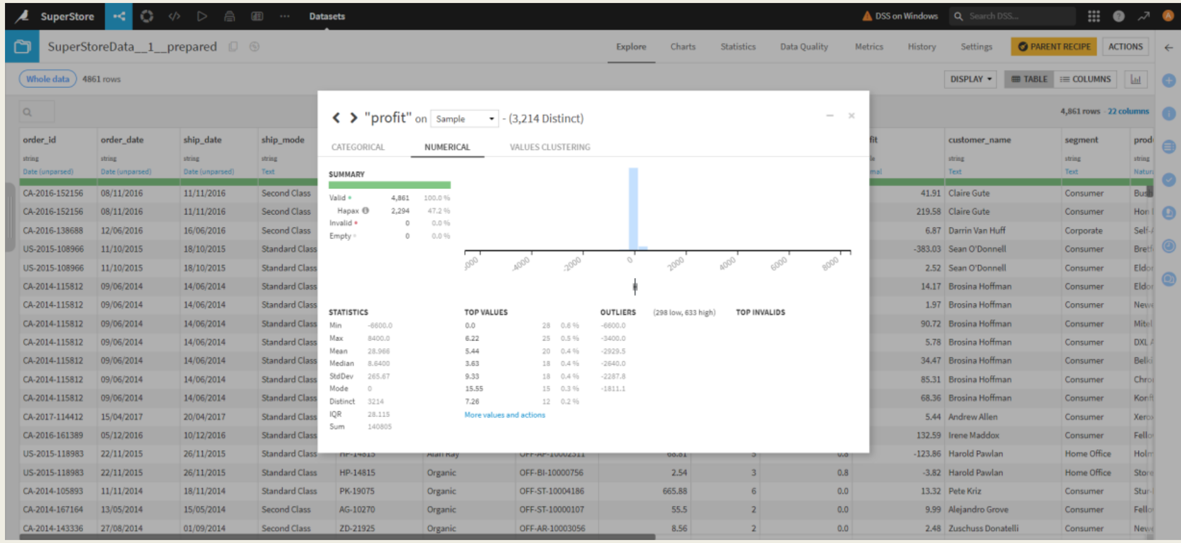
TOP VALUES

12.96	57	0.6 %
15.55	39	0.4 %
19.44	39	0.4 %
10.37	36	0.4 %
25.92	36	0.4 %
32.4	28	0.3 %
17.94	21	0.2 %

More values and actions

OUTLIERS (0 low, 1167 high): 499.17, 499.58, 499.95, 499.98, 499.99, 500.24

TOP INVALIDS



SuperStoreData__1__prepared — Whole data 4861 rows

"profit" on Sample - (3,214 Distinct)

CATEGORICAL | NUMERICAL | VALUES CLUSTERING

SUMMARY

Valid	4,861	100.0 %
Hapax	2,294	47.2 %
Invalid	0	0.0 %
Empty	0	0.0 %

STATISTICS

Min	-6600.0
Max	8400.0
Mean	28.966
Median	8.6400
StdDev	265.67
Mode	0
Distinct	3214
IQR	28.115
Sum	140805

TOP VALUES

0.0	28	0.6 %
6.22	25	0.5 %
5.44	20	0.4 %
3.63	18	0.4 %
9.33	18	0.4 %
15.55	15	0.3 %
7.26	12	0.2 %

More values and actions

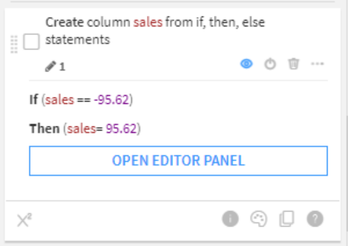
OUTLIERS (298 low, 633 high): -6600.0, -3400.0, -2929.5, -2640.0, -2287.8, -1811.1

TOP INVALIDS

→ Il y a un doublon, on le supprime rapidement.

→ Il y a une valeur aberrante dans « sales »: le « -95,62 ».

→ *On constate qu'il s'agit d'une erreur de frappe donc on enlève le signe « - ».*



□ Create column sales from if, then, else statements

If (sales == -95.62)

Then (sales= 95.62)

OPEN EDITOR PANEL

2. EDA

Dataiku