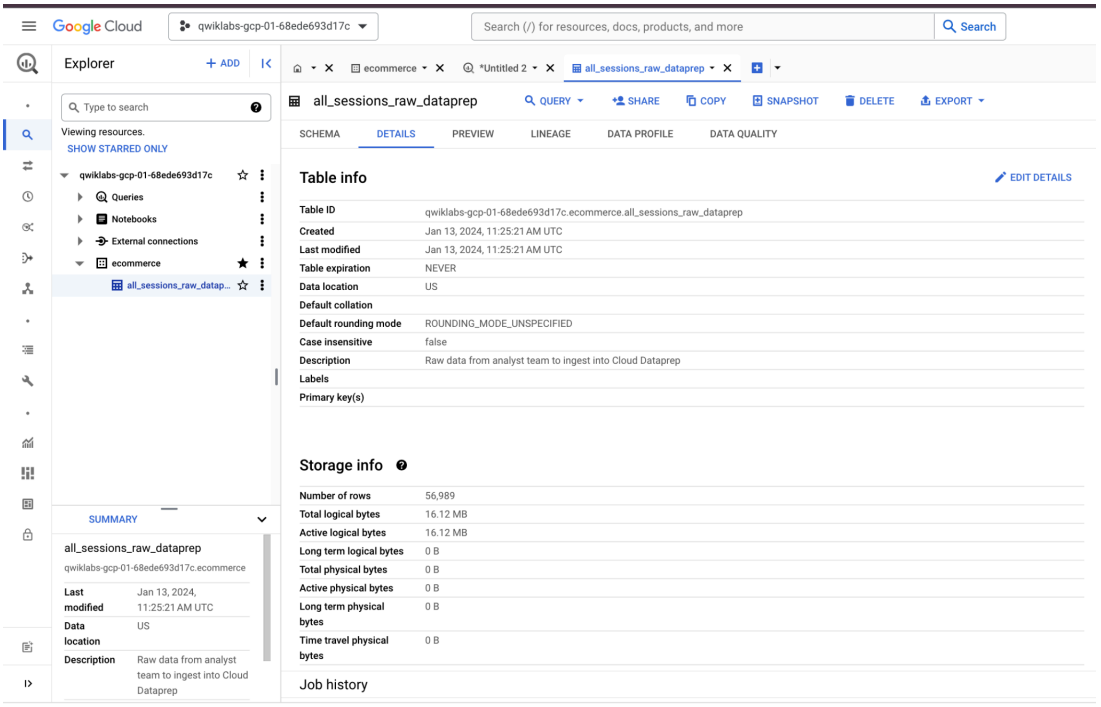
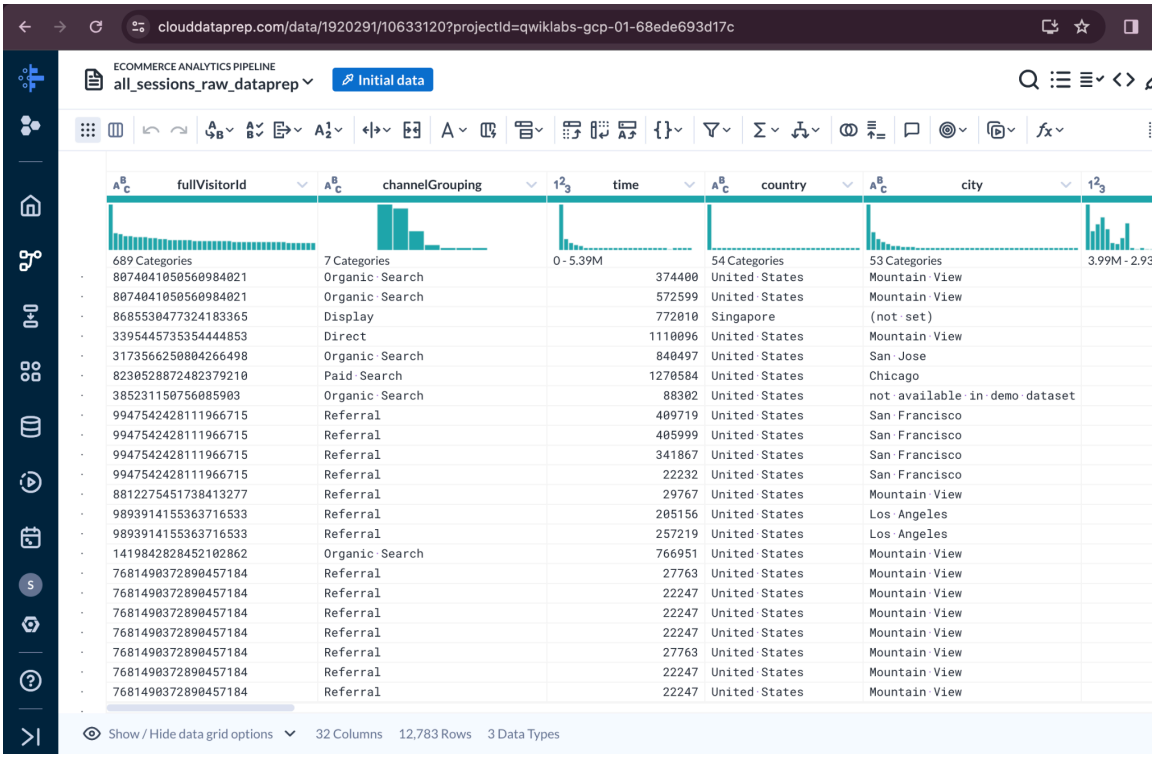
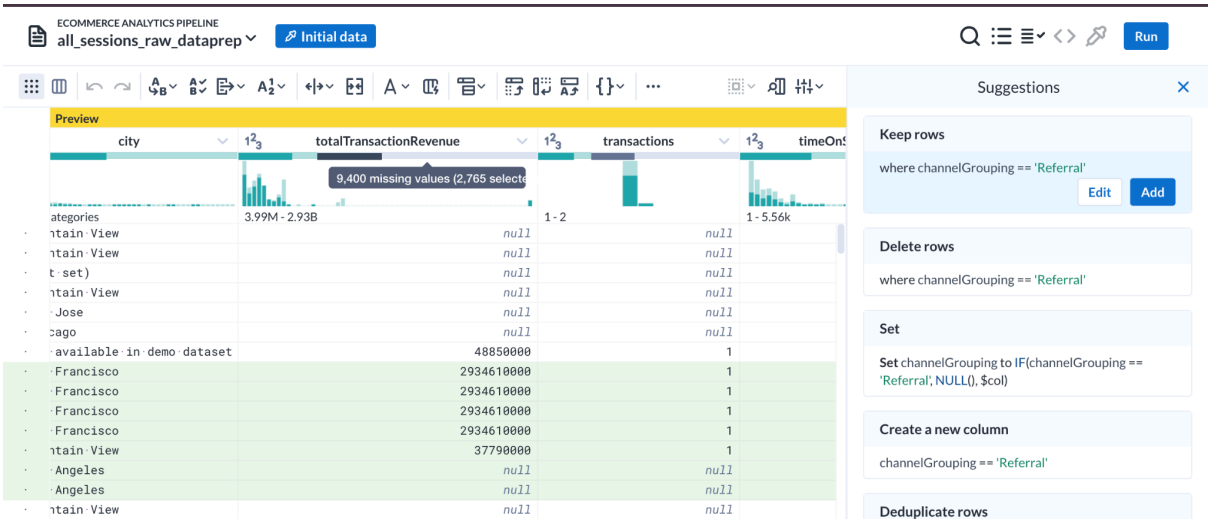
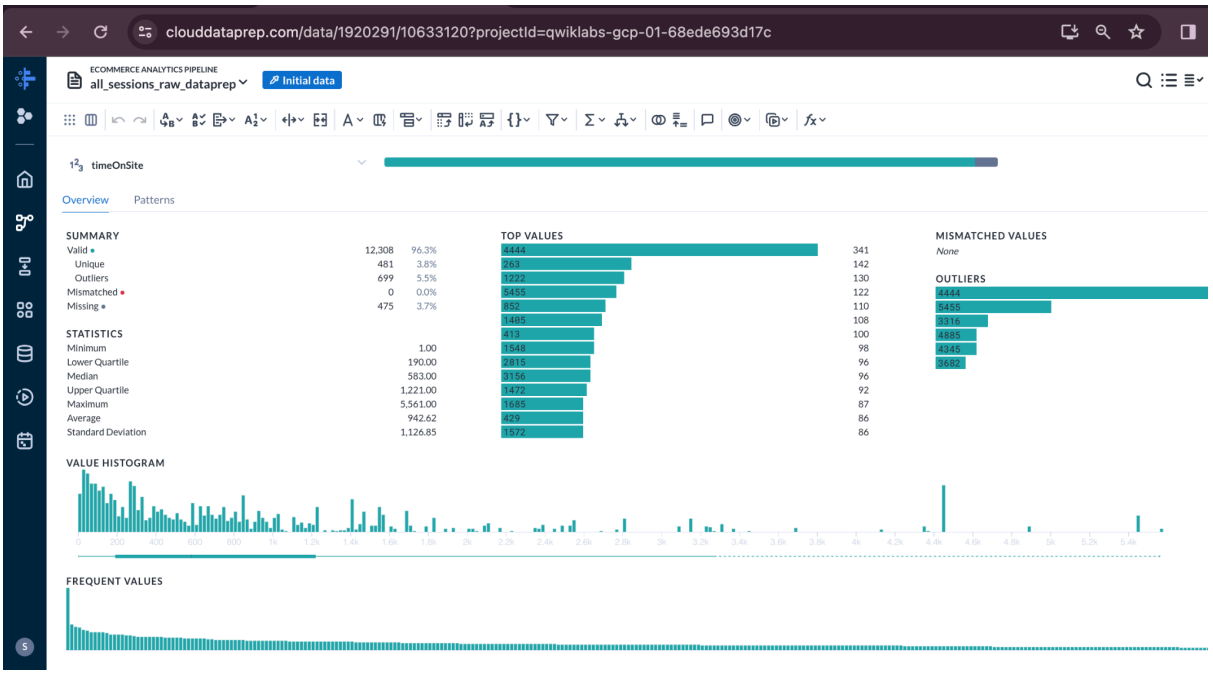
Connecting BigQuery data to Cloud Dataprep

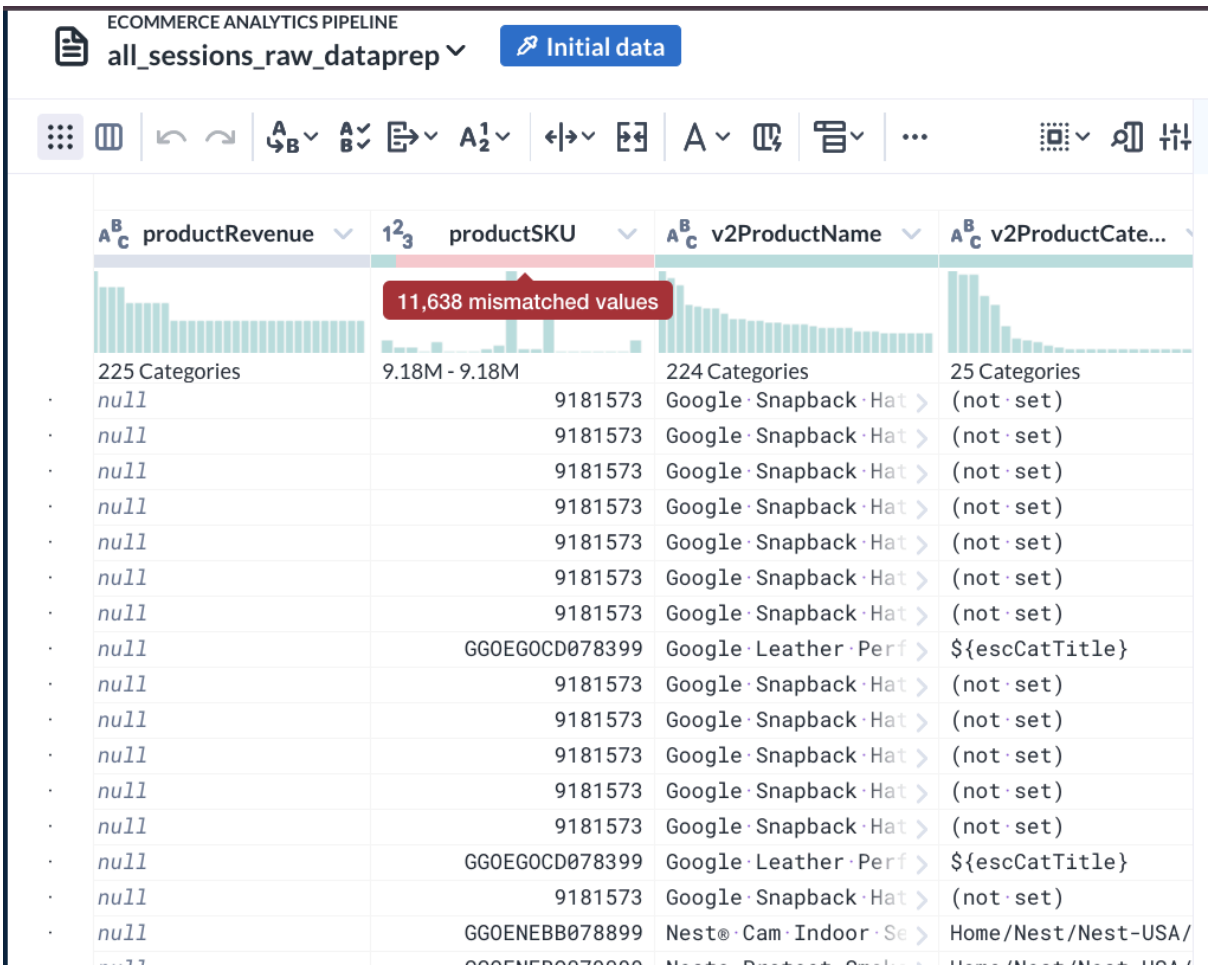


Exploring ecommerce data fields with the UI Alteryx




ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
Run
Preview
city
totalTransactionRevenue
transactions
timeOnS
9,400 missing values (2,765 selecte
3.99M - 2.93B
1 - 2
1 - 5.56k
Suggestions
Keep rows
where channelGrouping == 'Referral'
Edit
Add
Delete rows
where channelGrouping == 'Referral'
Set
Set channelGrouping to IF(channelGrouping == 'Referral', NULL(), $col)
Create a new column
channelGrouping == 'Referral'
Deduplicate rows


clouddataprep.com/data/1920291/10633120?projectId=qwiklabs-gcp-01-68ede693d17c
ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
timeOnSite
Overview
Patterns
SUMMARY
STATISTICS
TOP VALUES
MISMATCHED VALUES
OUTLIERS
VALUE HISTOGRAM
FREQUENT VALUES

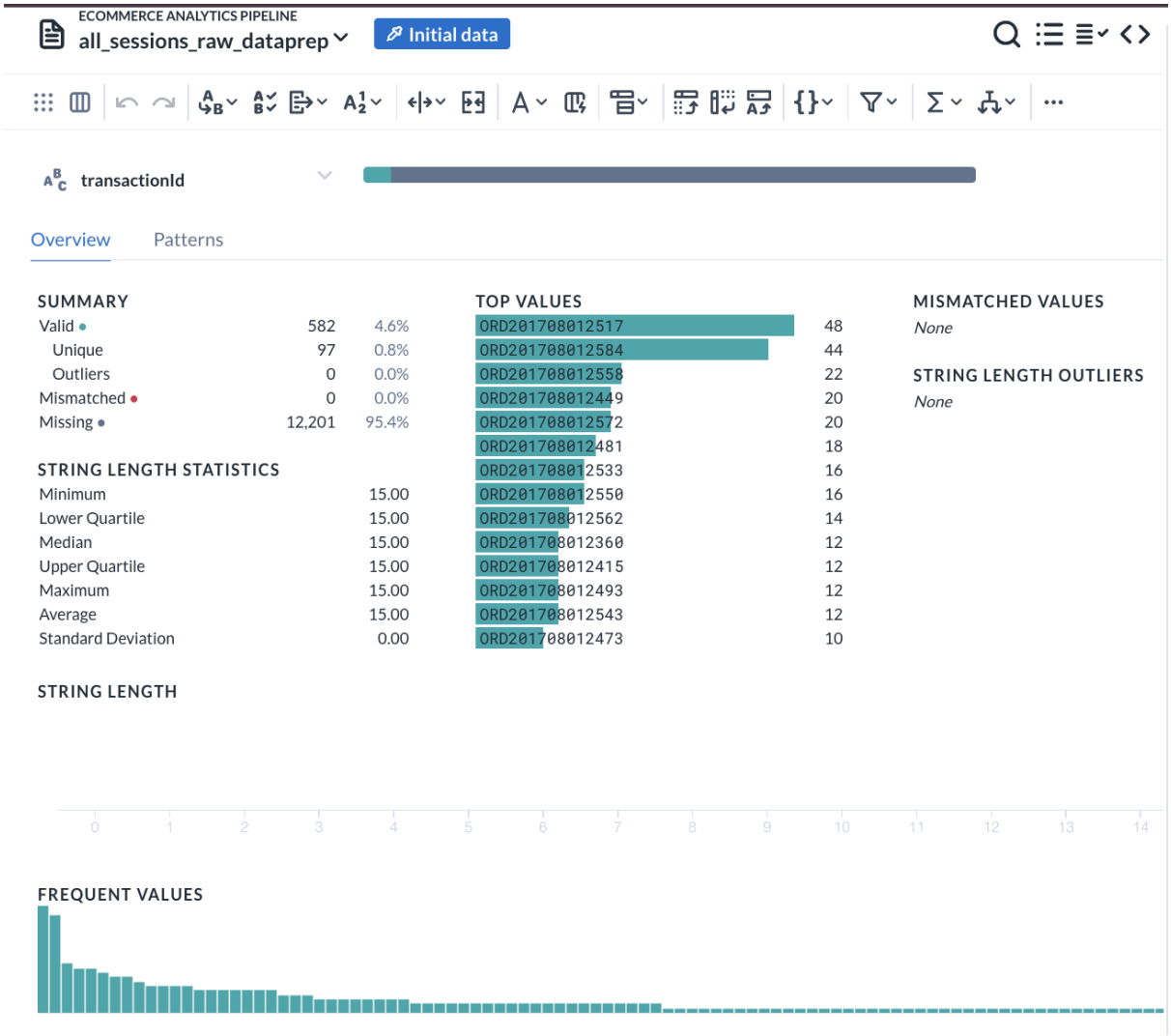

ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
productRevenue
productSKU
v2ProductName
v2ProductCate...
11,638 mismatched values
225 Categories
9.18M - 9.18M
224 Categories
25 Categories
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null GGOEGOCD078399 Google Leather Perf ${escCatTitle}
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null 9181573 Google Snapback Hat (not set)
null GGOEGOCD078399 Google Leather Perf ${escCatTitle}
null 9181573 Google Snapback Hat (not set)
null GGOENEBB078899 Nest® Cam Indoor Se Home/Nest/Nest-USA/

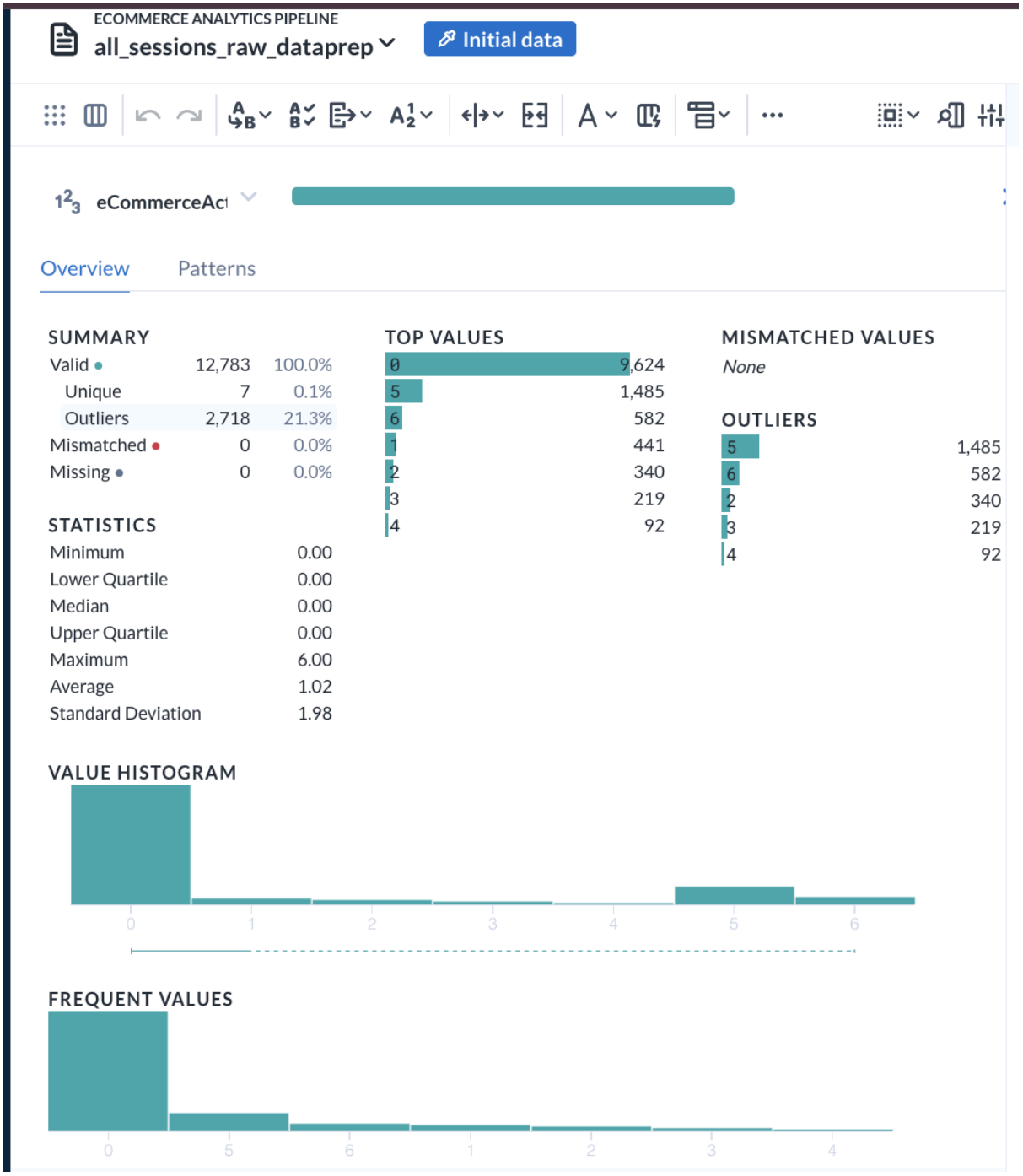
ECOMMERCE ANALYTICS PIPELINE

all_sessions_raw_dataprep

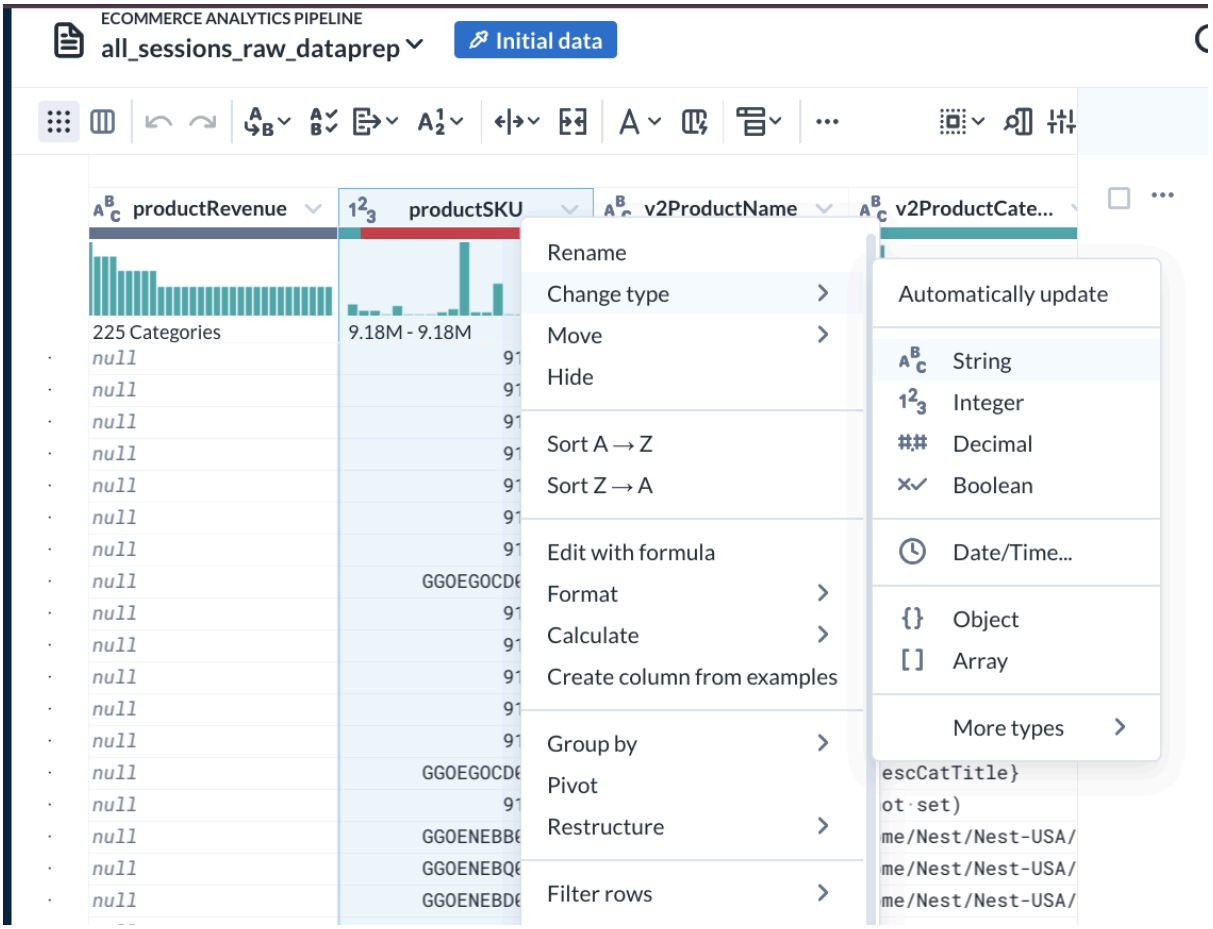
Initial data

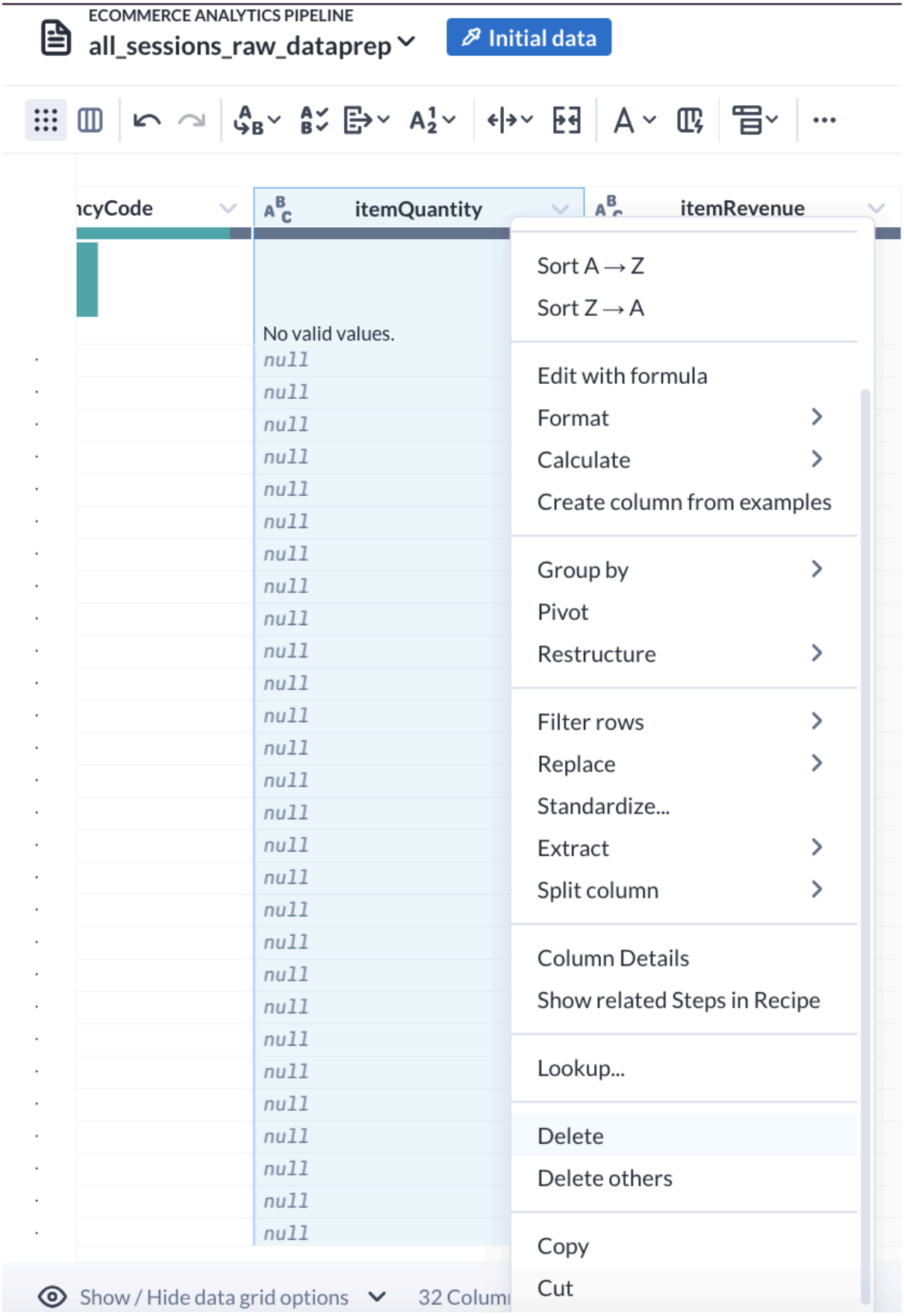
	itemQuantity	itemRevenue	transactionRev...	transactio
	No valid values.	No valid values.	89 Categories	97 Categories
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null
	null	null	null	null

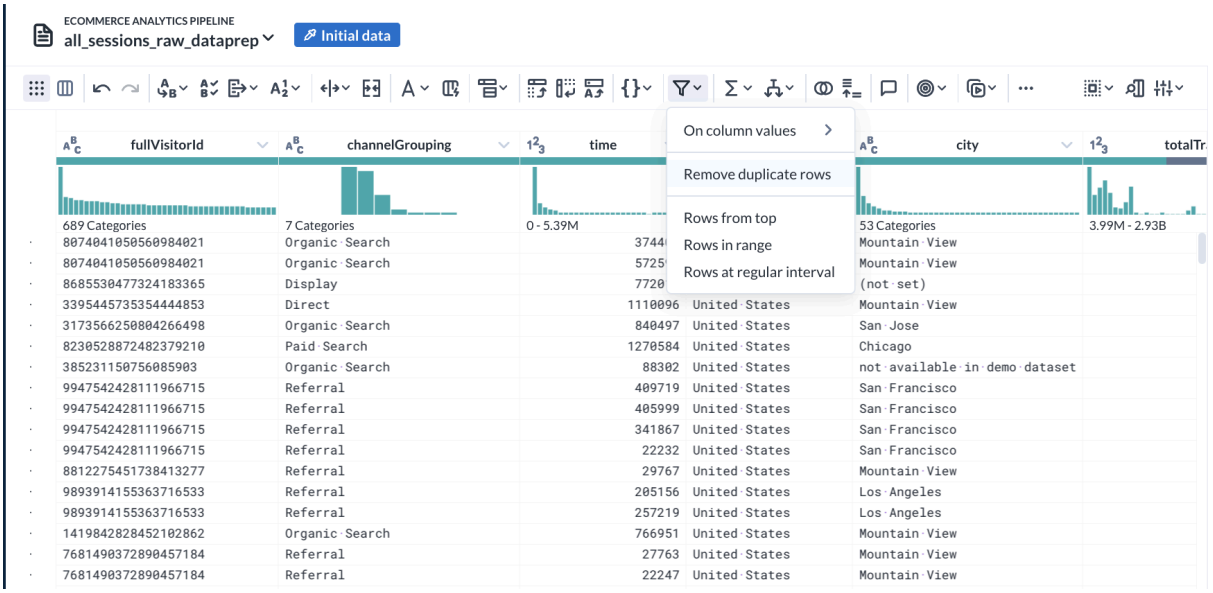

ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
transactionId
Overview
Patterns
SUMMARY
Valid 582 4.6%
Unique 97 0.8%
Outliers 0 0.0%
Mismatched 0 0.0%
Missing 12,201 95.4%
STRING LENGTH STATISTICS
Minimum 15.00
Lower Quartile 15.00
Median 15.00
Upper Quartile 15.00
Maximum 15.00
Average 15.00
Standard Deviation 0.00
TOP VALUES
ORD201708012517 48
ORD201708012584 44
ORD201708012558 22
ORD201708012449 20
ORD201708012572 20
ORD201708012481 18
ORD201708012533 16
ORD201708012550 16
ORD201708012562 14
ORD201708012360 12
ORD201708012415 12
ORD201708012493 12
ORD201708012543 12
ORD201708012473 10
MISMATCHED VALUES
None
STRING LENGTH OUTLIERS
None
STRING LENGTH
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
FREQUENT VALUES



Cleaning the data


ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
productRevenue
productSKU
v2ProductName
v2ProductCate...
225 Categories
9.18M - 9.18M
Rename
Change type
Move
Hide
Sort A → Z
Sort Z → A
Edit with formula
Format
Calculate
Create column from examples
Group by
Pivot
Restructure
Filter rows
Automatically update
String
Integer
Decimal
Boolean
Date/Time...
Object
Array
More types


ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
ncyCode
itemQuantity
itemRevenue
No valid values.
null
Sort A → Z
Sort Z → A
Edit with formula
Format
Calculate
Create column from examples
Group by
Pivot
Restructure
Filter rows
Replace
Standardize...
Extract
Split column
Column Details
Show related Steps in Recipe
Lookup...
Delete
Delete others
Copy
Cut
Show / Hide data grid options
32 Colum



New Step

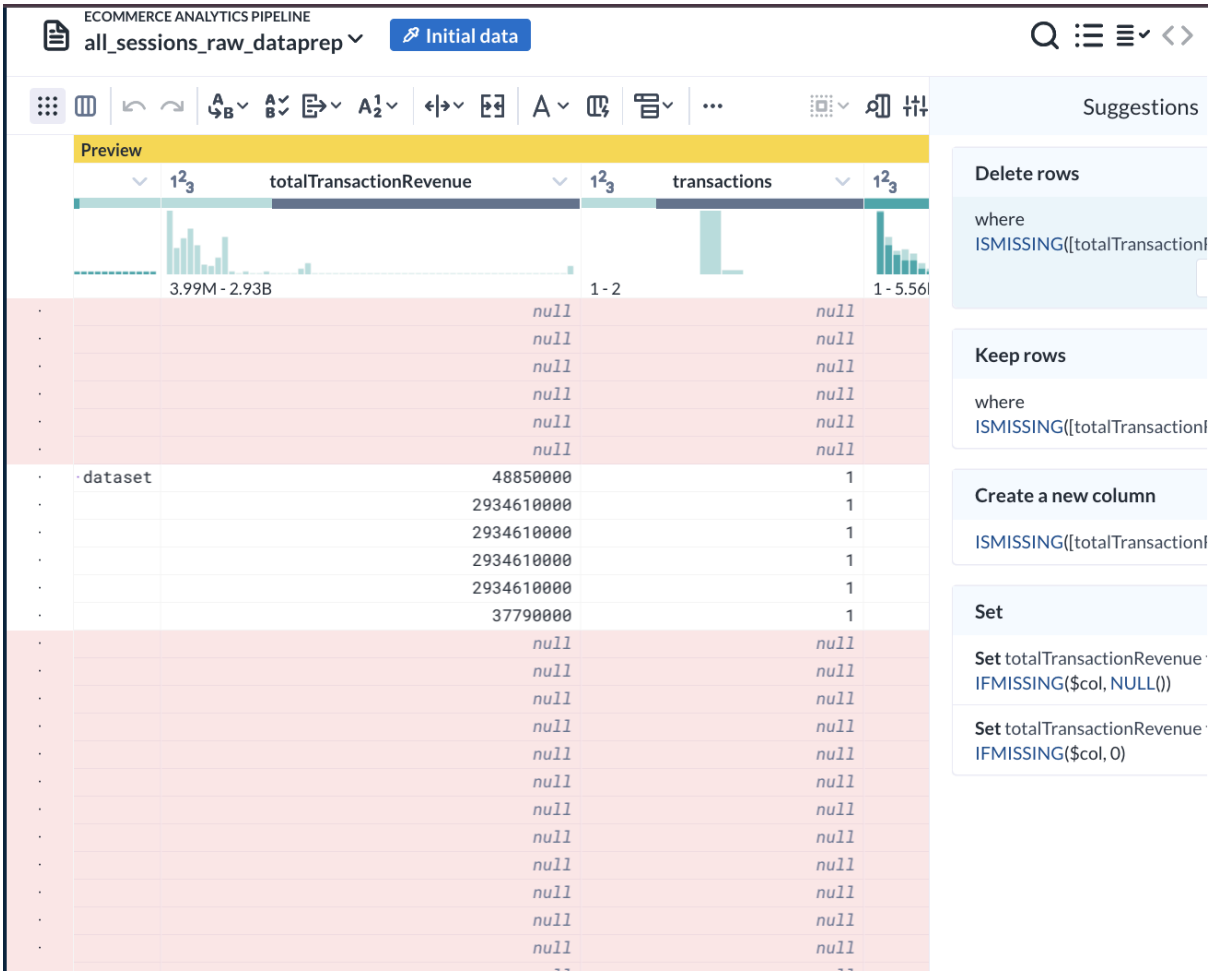
Recipe

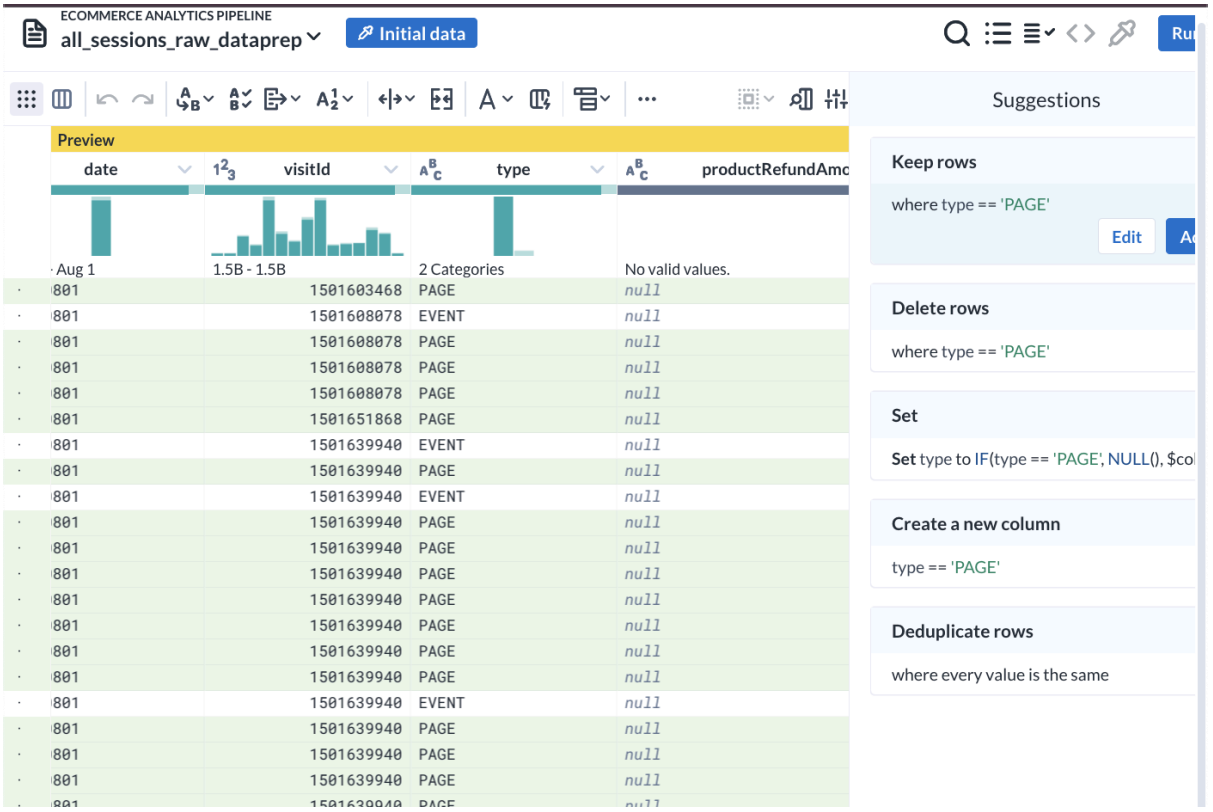
1 **Lock** productSKU **type** to String

2 **Delete** itemQuantity

3 **Delete** itemRevenue

4 **Remove duplicate** rows


ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
Suggestions
Preview
totalTransactionRevenue
transactions
3.99M - 2.93B
1 - 2
1 - 5.56
null
dataset
48850000
2934610000
37790000
1
Delete rows
where ISMISSING([totalTransactionI
Keep rows
where ISMISSING([totalTransactionI
Create a new column
ISMISSING([totalTransactionI
Set
Set totalTransactionRevenue
IFMISSING($col, NULL())
Set totalTransactionRevenue
IFMISSING($col, 0)


ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep
Initial data
Suggestions
Preview
date
visitId
type
productRefundAmo
Aug 1
1.5B - 1.5B
2 Categories
No valid values.
801
1501603468
1501608078
1501651868
1501639940
PAGE
EVENT
null
Keep rows
where type == 'PAGE'
Edit
Delete rows
where type == 'PAGE'
Set
Set type to IF(type == 'PAGE', NULL(), $co
Create a new column
type == 'PAGE'
Deduplicate rows
where every value is the same

New Step

Recipe

1 **Lock** productSKU **type** to String

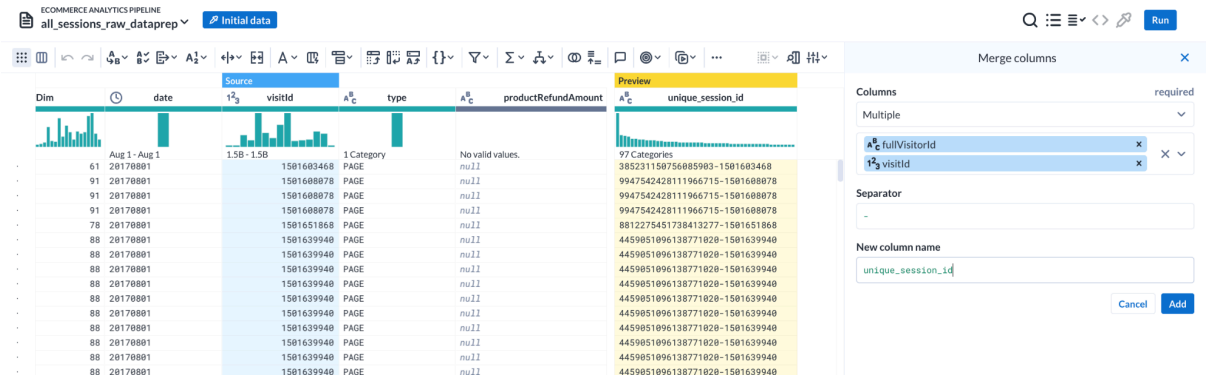
2 **Delete** itemQuantity

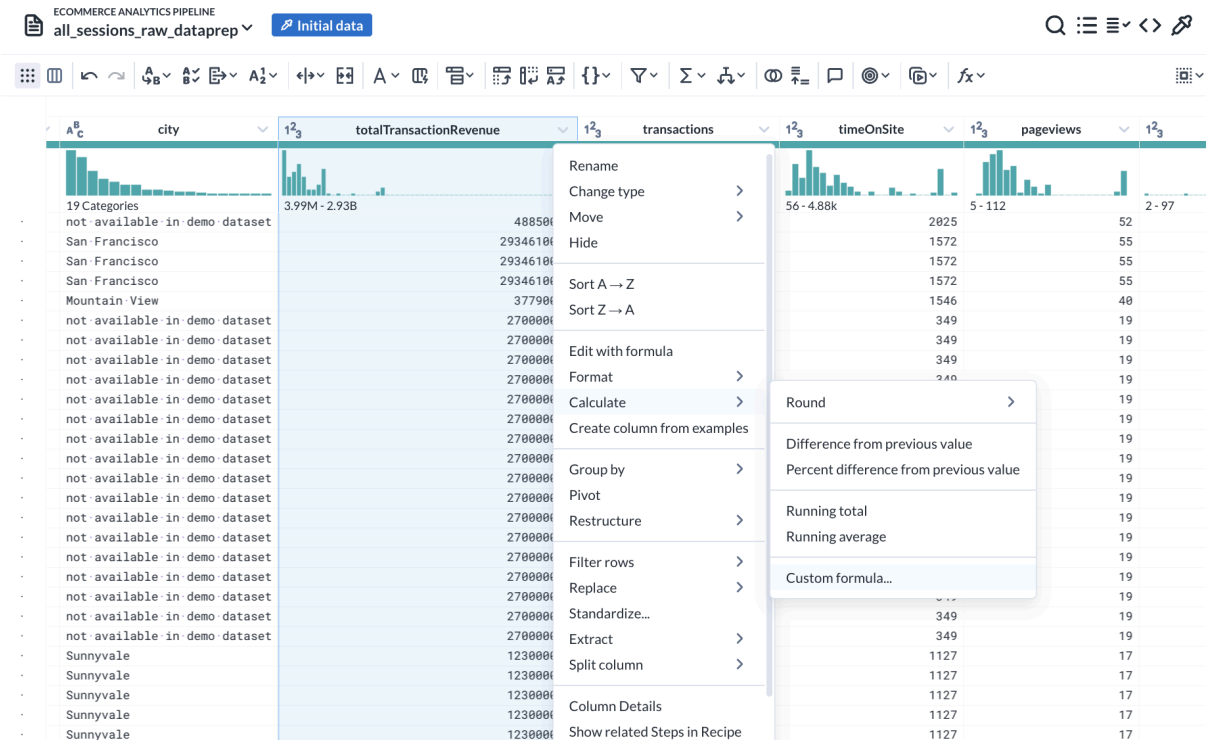
3 **Delete** itemRevenue

4 **Remove duplicate** rows

5 **Delete rows** where ISMISSING([totalTransactionRevenue])

6 **Keep rows** where type == 'PAGE'





Run Job

Running Environment

Trifacta Photon
Run job on Trifacta Photon (best for small and medium-sized jobs, up to approximately 1 GB of data)

✓ **Dataflow + BigQuery**
Run job with hybrid execution on Dataflow and BigQuery. Job will be fully executed in BigQuery if all transforms and functions are supported. Learn more

Options

☑ Profile results and assess data quality rules
Generate a statistical profile of the published data and evaluate data quality rules

☑ Validate schema
Analyze input schema and compare with the flow dataset schema.
Currently applicable to relational datasets and limited file types. Learn more.

☐ Fail job if dataset schemas change
Fail the job if the input data schema differs from the schema of the flow dataset.

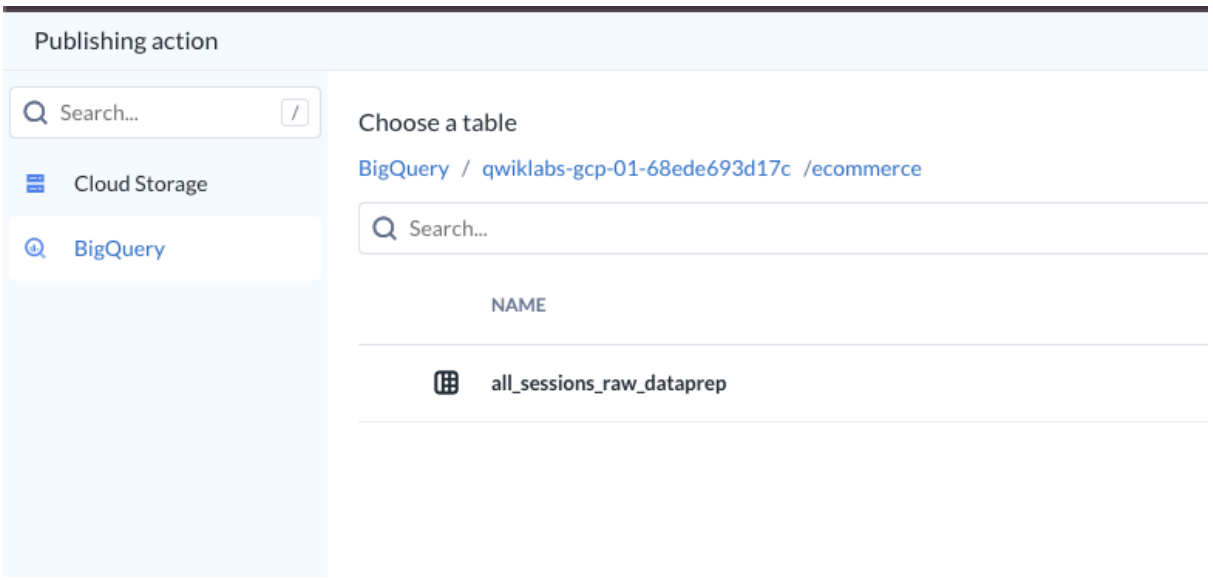
☐ Ignore recipe errors
Allow jobs to be run even if there are errors present in recipe steps

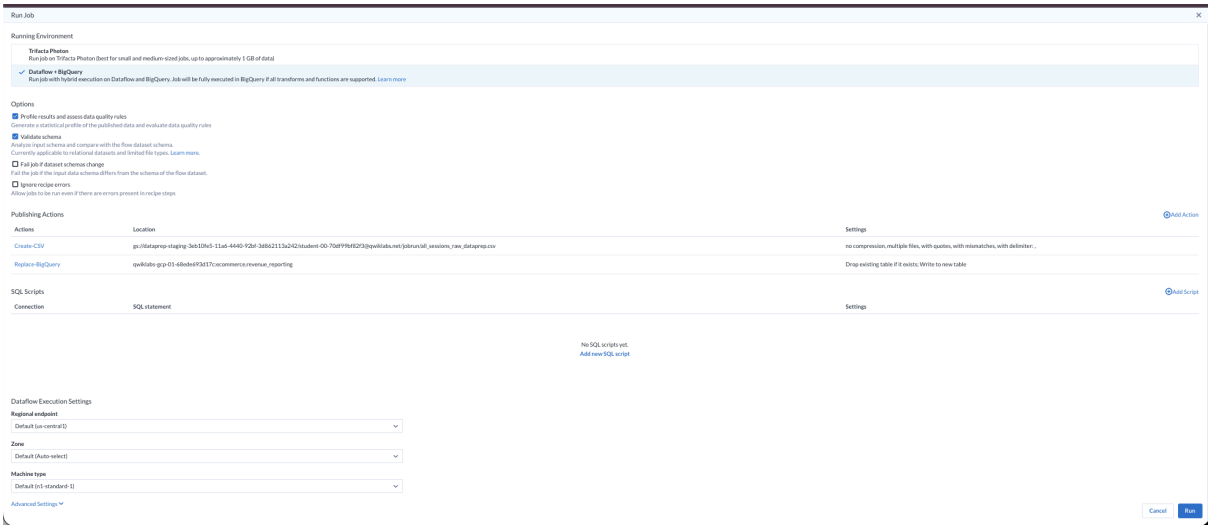
Publishing Actions

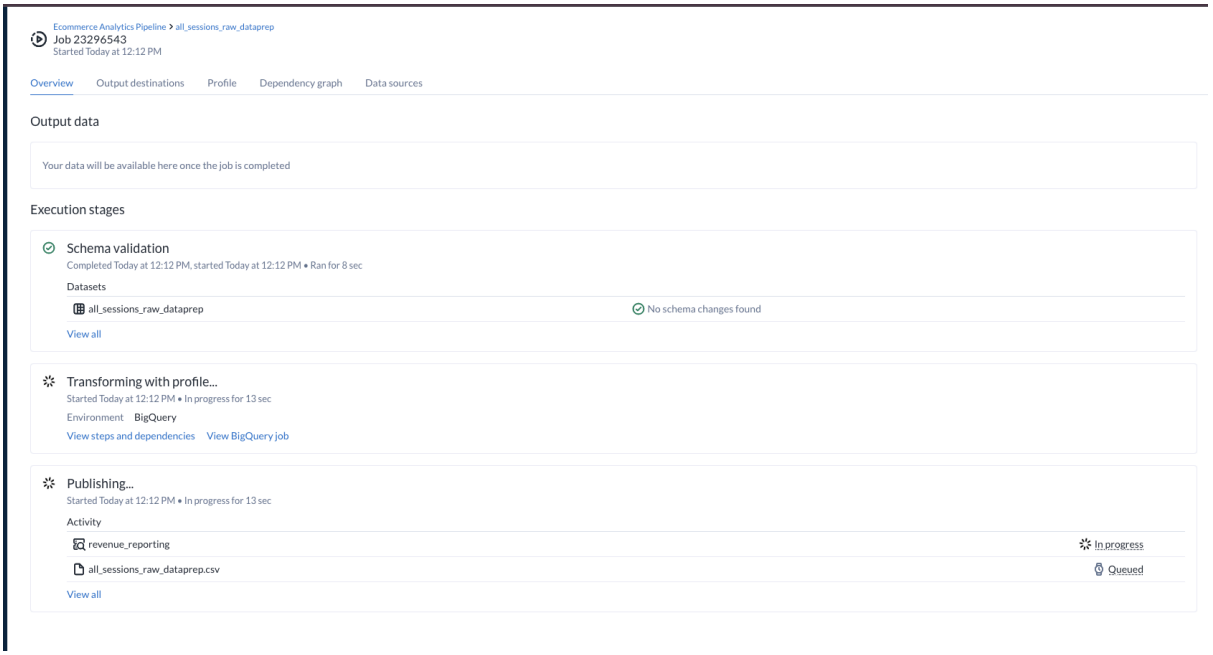
Actions	Location	Settings
Create-CSV	gs://dataprep-staging-3eb10fe5-11a6-4440-92bf-3d862113a242/student-00-70df99bf82f3@qwiklabs.net/jobrun/all_sessions_raw_dataprep.csv	no compression, multiple files, with quotes, with mismatches, with delimiter: ,

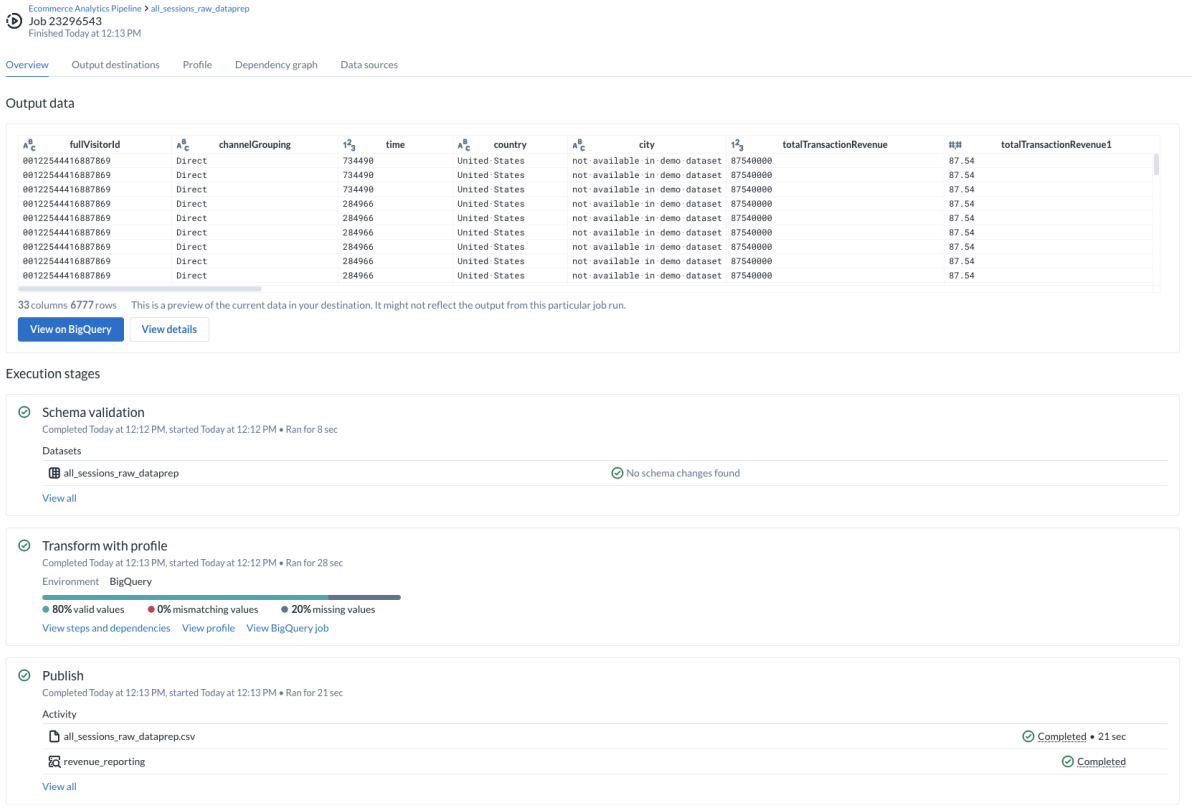
SQL Scripts

Connection	SQL statement	Settings


Publishing action
Search...
/
Cloud Storage
BigQuery
Choose a table
BigQuery / qwiklabs-gcp-01-68ede693d17c /ecommerce
Search...
NAME
all_sessions_raw_dataprep


Run Job
Running Environment
Trifacta Photon
Run job on Trifacta Photon (best for small and medium-sized jobs, up to approximately 1 GB of data)
Dataflow + BigQuery
Run job with hybrid execution on Dataflow and BigQuery. Job will be fully executed in BigQuery if all transforms and functions are supported. Learn more
Options
Profile results and assess data quality rules
Generate a statistical profile of the published data and evaluate data quality rules
Validate schema
Analyze input schema and compare with the flow dataset schema.
Currently applicable to relational datasets and limited file types. Learn more.
Fail job if dataset schemas change
Fail the job if the input data schema differs from the schema of the flow dataset.
Ignore recipe errors
Allow jobs to be run even if there are errors present in recipe steps
Publishing Actions
Add Action
Actions
Location
Settings
Create-CSV
gs://dataprep-staging-3eb10fe5-11a6-4440-92bf-3d862113a242/student-00-70df99bf82f3@qwiklabs.net/jobruns/all_sessions_raw_dataprep.csv
no compression, multiple files, with quotes, with mismatches, with delimiter: ,
Replace-BigQuery
qwiklabs-gcp-01-68ede693d17c:ecommerce.revenue_reporting
Drop existing table if it exists; Write to new table
SQL Scripts
Add Script
Connection
SQL statement
Settings
No SQL scripts yet.
Add new SQL script
Dataflow Execution Settings
Regional endpoint
Default (us-central1)
Zone
Default (Auto-select)
Machine type
Default (n1-standard-1)
Advanced Settings
Cancel
Run


Ecommerce Analytics Pipeline > all_sessions_raw_dataprep
Job 23296543
Started Today at 12:12 PM
Overview
Output destinations
Profile
Dependency graph
Data sources
Output data
Your data will be available here once the job is completed
Execution stages
Schema validation
Completed Today at 12:12 PM, started Today at 12:12 PM • Ran for 8 sec
Datasets
all_sessions_raw_dataprep
No schema changes found
View all
Transforming with profile...
Started Today at 12:12 PM • In progress for 13 sec
Environment BigQuery
View steps and dependencies
View BigQuery job
Publishing...
Started Today at 12:12 PM • In progress for 13 sec
Activity
revenue_reporting
In progress
all_sessions_raw_dataprep.csv
Queued
View all


Ecommerce Analytics Pipeline > all_sessions_raw_dataprep
Job 23296543
Finished Today at 12:13 PM
Overview
Output destinations
Profile
Dependency graph
Data sources
Output data
fullVisitorId
channelGrouping
time
country
city
totalTransactionRevenue
totalTransactionRevenue1
0012254441688786 9 Direct 734490 United States not available in demo dataset 87540000 87.54
33 columns 6777 rows
This is a preview of the current data in your destination. It might not reflect the output from this particular job run.
View on BigQuery
View details
Execution stages
Schema validation
Completed Today at 12:12 PM, started Today at 12:12 PM • Ran for 8 sec
Datasets
all_sessions_raw_dataprep
No schema changes found
View all
Transform with profile
Completed Today at 12:13 PM, started Today at 12:12 PM • Ran for 28 sec
Environment BigQuery
80% valid values
0% mismatching values
20% missing values
View steps and dependencies
View profile
View BigQuery job
Publish
Completed Today at 12:13 PM, started Today at 12:13 PM • Ran for 21 sec
Activity
all_sessions_raw_dataprep.csv
Completed • 21 sec
revenue_reporting
Completed
View all