Module 5

Cleaning and Transforming your Data

In this module we will:

- **Examine the 5 Principles of Dataset Integrity**
- Characterize Dataset Shape and Skew
- Clean and Transform Data using SQL
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

Garbage in… garbage out

High quality datasets conform to strict integrity rules

1

Validity

Data conforms to your business rules

⚠

Challenges

Out of Range

Empty Fields

Data Mismatch

2

Accuracy

Data conforms to an objective true value.

⚠

Challenges

Lookup Datasets Do Not Exist

3

Completeness

Create, save, and store datasets.

⚠

Challenges

Missing Data

4

Consistency

Derive insights from data.

⚠

Challenges

Duplicate Records

Concurrency Issues

5

Uniformity

Explore and present data

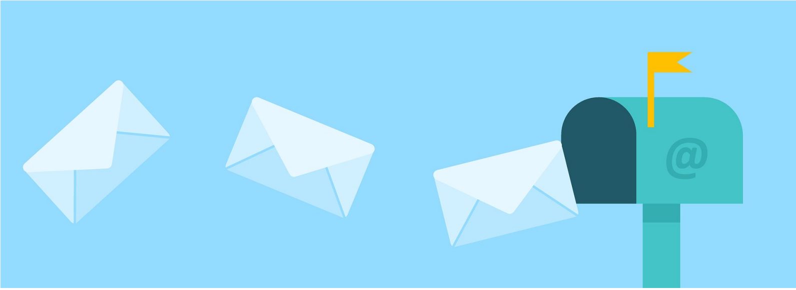
⚠

Challenges

Same Units of Measurement

Google Cloud

Valid data follows constraints on uniqueness



what do these identifiers have in common? why were they set up that way?

Valid data corresponds to range constraints

Roll #	Value
1	2
2	2
3	6
4	5
5	1
6	7

which value(s) are out of range?

Accurate data matches to a known source of truth

Washington
Montana
North Dakota
Minnesota
Wisconsin
Michigan
New Hampshire
Vermont
Massachusetts
Maine
Oregon
Idaho
Wyoming
South Dakota
New York
Rhode Island
Connecticut
Nevada
Utah
Nebraska
Iowa
Pennsylvania
New Jersey
Illinois
Indiana
Ohio
Delaware
Colorado
Kansas
Missouri
Maryland
California
Kentucky
Virginia
West Virginia
North Carolina
Oklahoma
Tennessee
Arizona
New Mexico
Arkansas
South Carolina
Alabama
Georgia
Mississippi
Texas
Louisiana
Alaska
Hawaii
Florida

U.S. States
Washington
Oregon
California
Hot Dog
Florida
Maine

Lamps and Clocks?

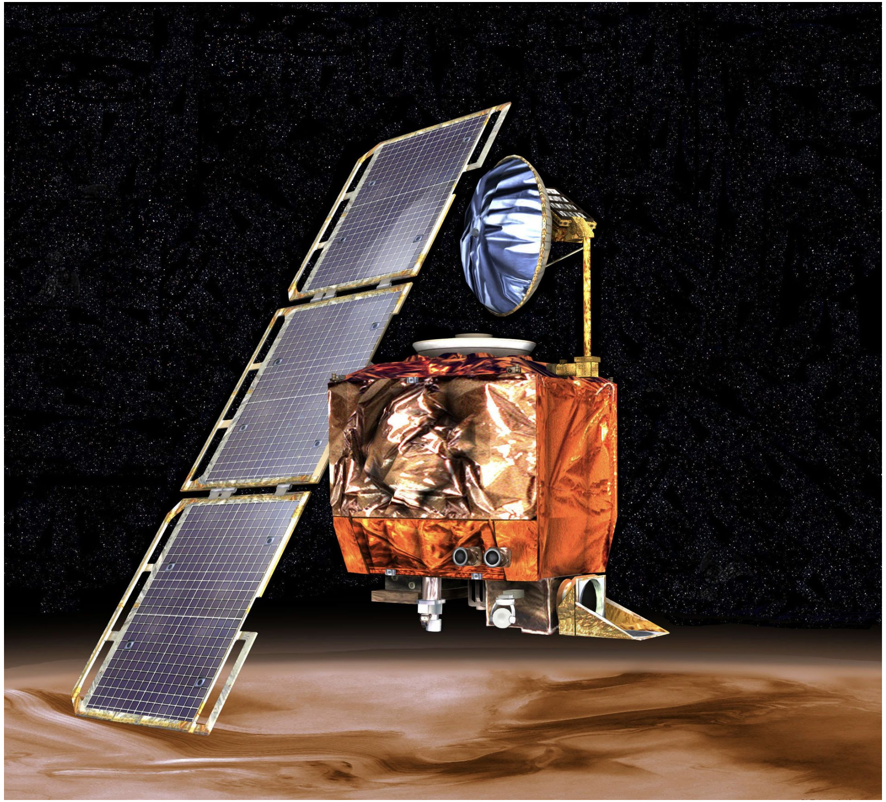
Consistent Data Ensures Harmony across Systems

House Address	Owner ID
123 ABC St	12

Owner ID	Owner Address
15	123 ABC St.
12	53rd Ave.

Who owns the house?

Uniformity in Data Means Measuring the Same Way



=$125 Million

In November 1999, NASA lost a Mars climate orbiter because of English vs Metric system measurements

Google Cloud

Module 5

Cleaning and Transforming your Data

In this module we will:

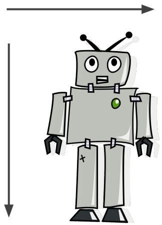
- Examine the 5 Principles of Dataset Integrity
- **Characterize Dataset Shape and Skew**
- Clean and Transform Data using SQL
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

Lab: Explore and Shape data with Cloud Dataprep

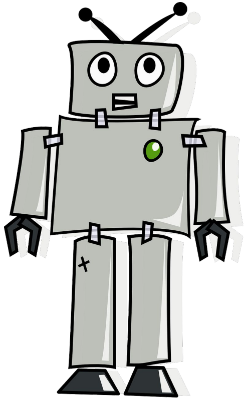
Understanding Dataset Shape

Number of Columns

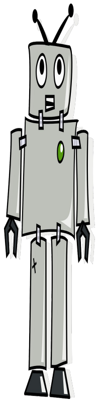
Number of Rows



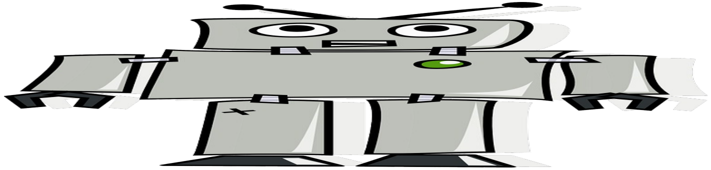
Small Dataset



Even Height and Width

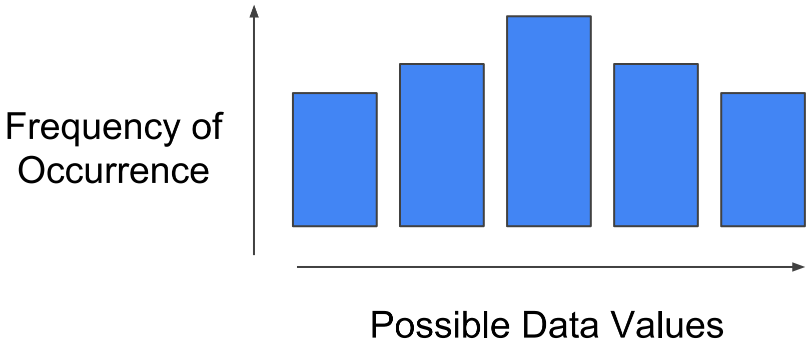


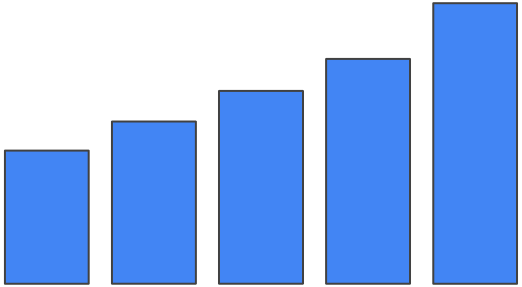
Taller than Wide



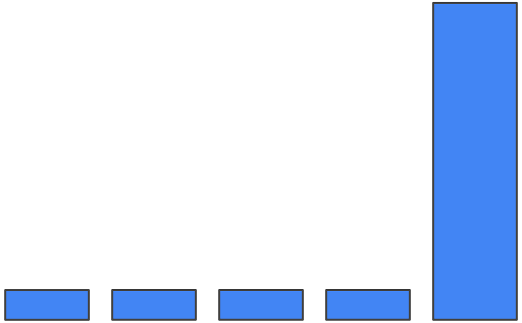
Wide but Short

Understanding Dataset Skew (Distribution of Values)

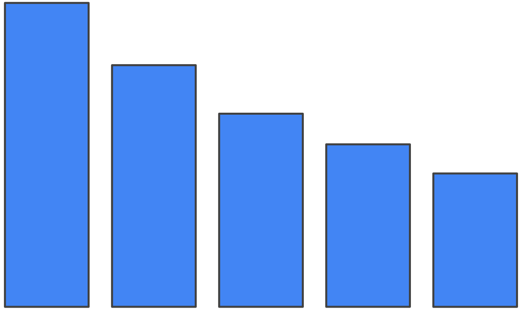




Skewed Right



Heavy Skew



Skewed Left

Module 5

Cleaning and Transforming your Data

In this module we will:

- Examine the 5 Principles of Dataset Integrity
- Characterize Dataset Shape and Skew
- **Clean and Transform Data using SQL**
- Clean and Transform Data using a new UI: Introducing Cloud Dataprep

Clean and Transform Data with SQL

1

Validity

Data conforms to your business rules

⚠

Challenges

Out of Range

Empty Fields

Data Mismatch

- Setup Field Data Type Constraints
- Specify fields as NULLABLE or REQUIRED
- Proactively check for NULL values
- Check and Filter for Allowable Range values
  - SQL Conditionals: CASE WHEN, IF ( )
- Require Primary Keys / Relational Constraints in upstream source systems (remember, BigQuery is an analytics warehouse not your primary operational database)

Google Cloud

Clean and Transform Data with SQL

2

Accuracy

Data conforms to an objective true value.

Challenges

Lookup Datasets Do Not Exist

- Create test cases or calculated fields to check values
  - SQL: (quantity_ordered * item_price) AS sub_total
- Lookup values against an objective reference dataset
  - SQL: IN( ) with a subquery or JOIN

Clean and Transform Data with SQL

3

Completeness

Create, save, and store datasets.

⚠

Challenges

Missing Data

- Thoroughly explore the existing dataset shape and skew and look for missing values
  - SQL: NULLIF( ) , IFNULL( ), COALESCE( )
- Enrich the existing dataset with others using UNIONs and JOINs
  - SQL: UNION, JOIN
  - Example: Multiple years of historical data are available for analysis

Clean and Transform Data with SQL

4

Consistency

Derive insights from data.

⚠

Challenges

Duplicate Records

Concurrency Issues

- Store one fact in one place and use IDs to lookup
- Use String Functions to clean data
  - PARSE_DATE( )
  - SUBSTR( )
  - REPLACE( )

Clean and Transform Data with SQL

5

Uniformity

Explore and present data

⚠

Challenges

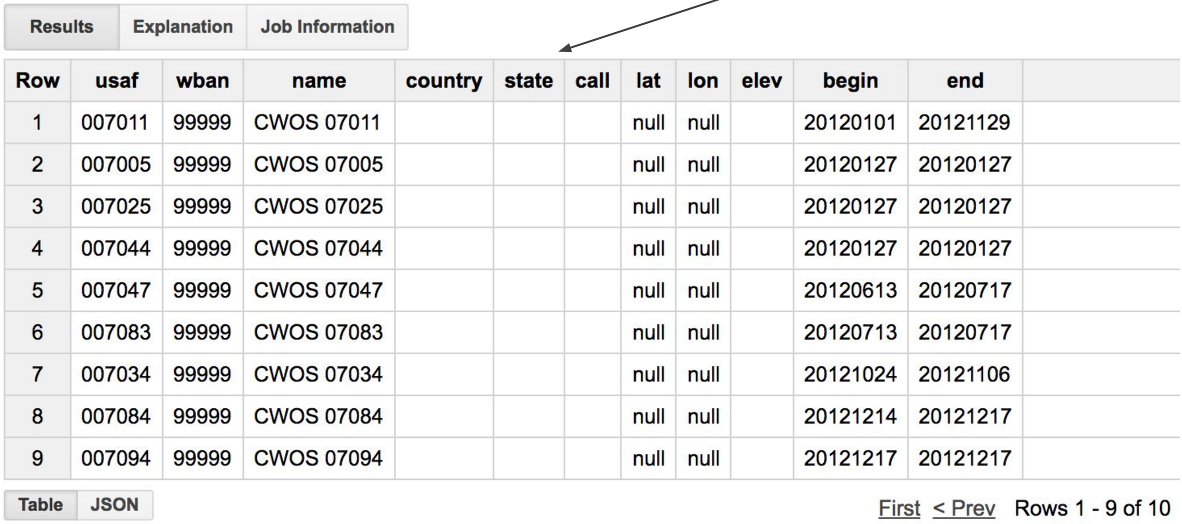
Same Units of Measurement

- Document and comment your approach
- Use FORMAT ( ) to clearly indicate units
- CAST( ) data types to the same format and digits
- Label all visualizations appropriately

Tricky NULLs when Filtering Out Missing Values

```
#standardSQL
SELECT * FROM
`bigquery-public-data.noaa_gsod.stations`
WHERE state IS NOT NULL
LIMIT 10
```

Why does the below query still show blank state values when we clearly filtered on IS NOT NULL?



Row	usaf	wban	name	country	state	call	lat	lon	elev	begin	end
1	007011	99999	CWOS 07011				null	null		20120101	20121129
2	007005	99999	CWOS 07005				null	null		20120127	20120127
3	007025	99999	CWOS 07025				null	null		20120127	20120127
4	007044	99999	CWOS 07044				null	null		20120127	20120127
5	007047	99999	CWOS 07047				null	null		20120613	20120717
6	007083	99999	CWOS 07083				null	null		20120713	20120717
7	007034	99999	CWOS 07034				null	null		20121024	20121106
8	007084	99999	CWOS 07084				null	null		20121214	20121217
9	007094	99999	CWOS 07094				null	null		20121217	20121217

Module 5

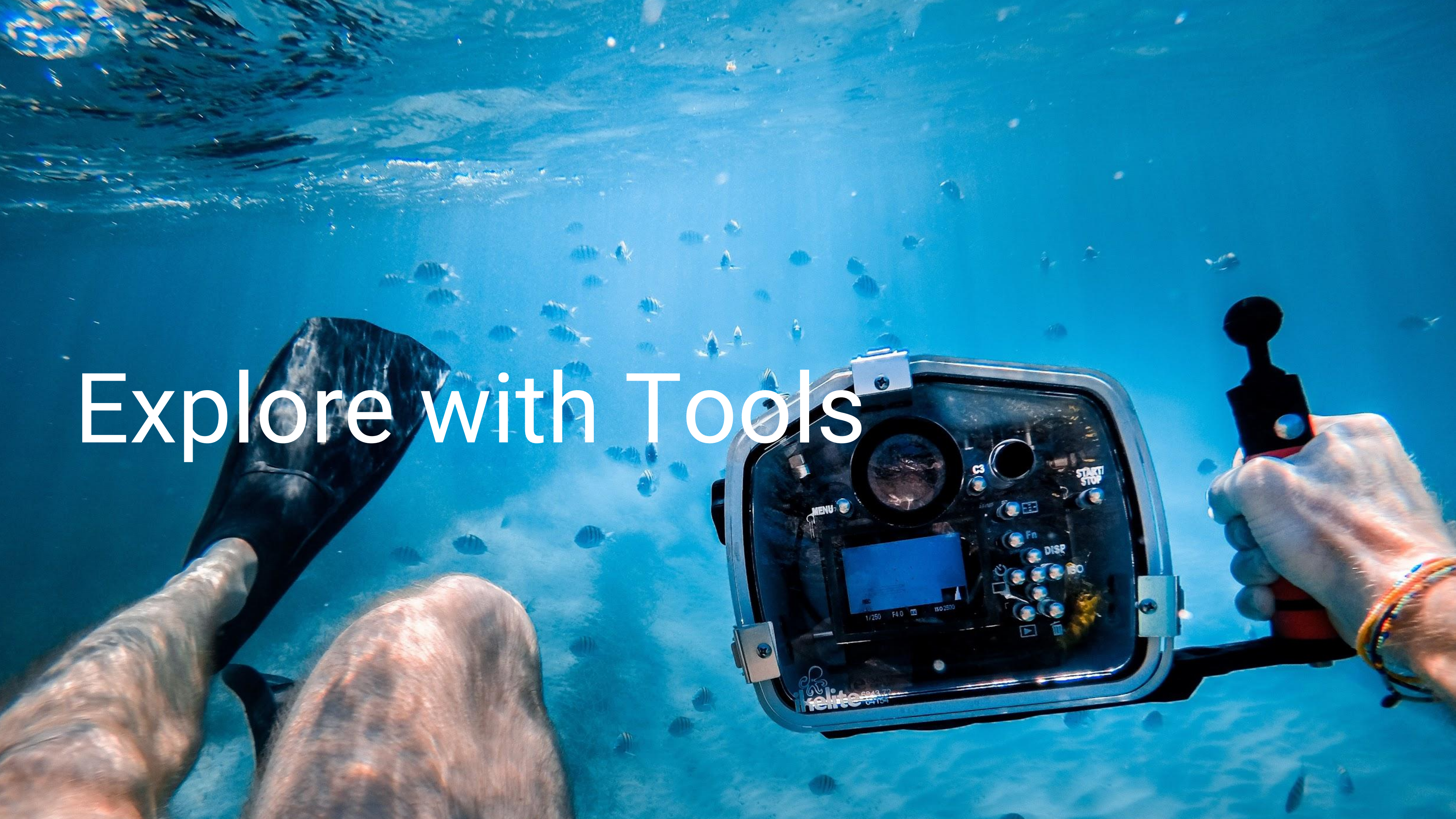
Cleaning and Transforming your Data

In this module we will:

- Examine the 5 Principles of Dataset Integrity
- Characterize Dataset Shape and Skew
- Clean and Transform Data using SQL
- **Clean and Transform Data using a new UI: Introducing Cloud Dataprep**

Google Cloud

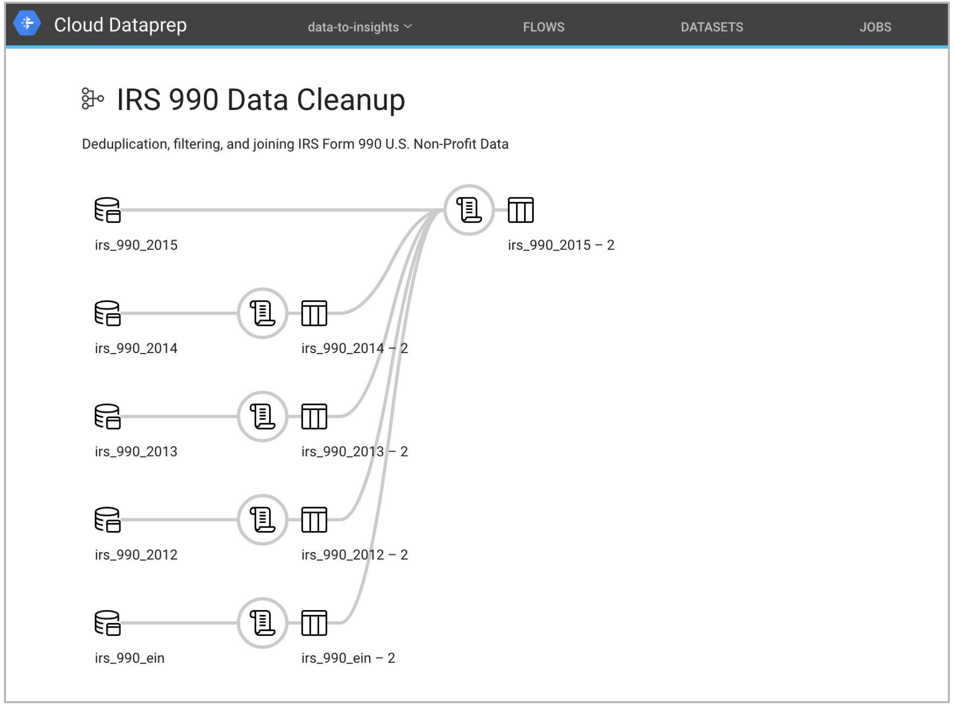
Explore with Tools



Create Repeatable Data **Transformation Flows in a UI**

Use Flows to wrangle your data.

Create Flow

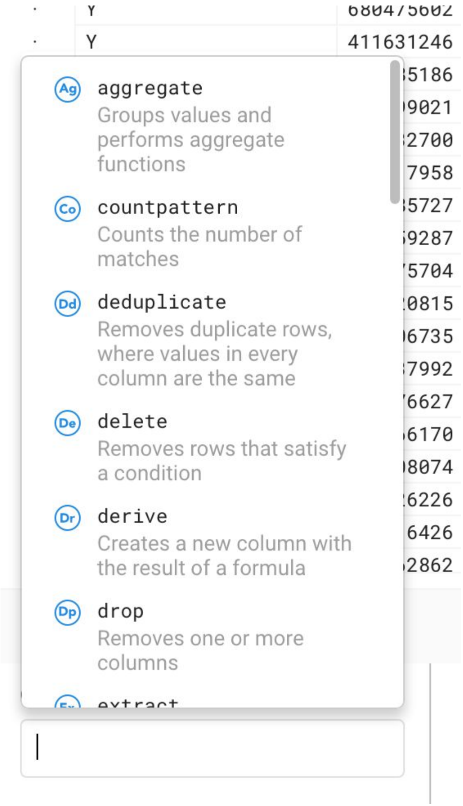


IRS 990 Data Cleanup

Deduplication, filtering, and joining IRS Form 990 U.S. Non-Profit Data

Transform Data with a Variety of **Predefined Wranglers**

- Use the Cloud Dataprep GUI to create and preview data preparation steps
- Chain together multiple wranglers into a repeatable recipe
- Common tasks like record deduplication and derived fields



Y 411631246

`aggregate`
Groups values and performs aggregate functions

`countpattern`
Counts the number of matches

`deduplicate`
Removes duplicate rows, where values in every column are the same

`delete`
Removes rows that satisfy a condition

`derive`
Creates a new column with the result of a formula

`drop`
Removes one or more columns

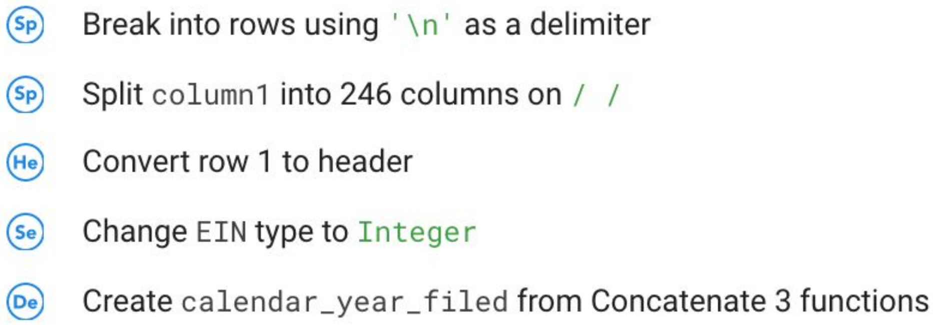
`extract`

Chain Transformation Rules Together into a **Recipe**

- Repeatable set of transformation steps build by chaining data wranglers together

- Jobs run against recipes

- Can include end-to-end steps from ingestion, transformation, aggregation, save to BigQuery



Sp Break into rows using `'\n'` as a delimiter

Sp Split `column1` into 246 columns on `/ /`

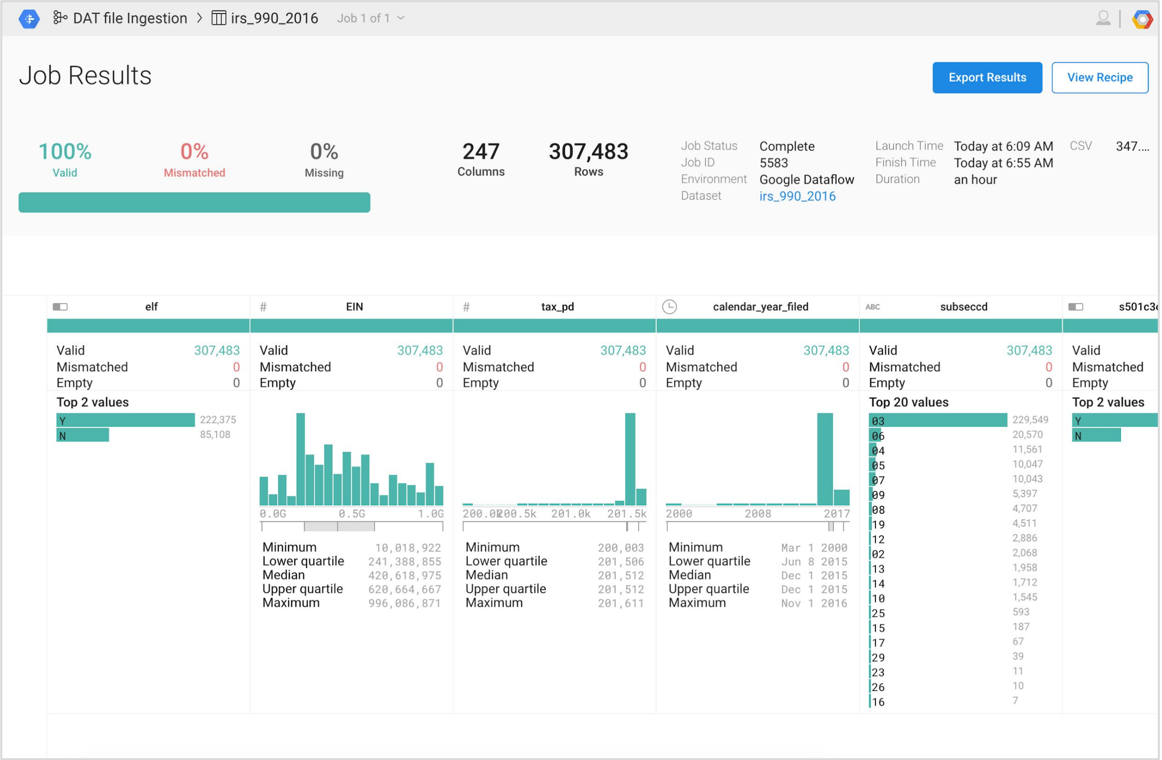
He Convert row 1 to header

Se Change `EIN` type to `Integer`

De Create `calendar_year_filed` from Concatenate 3 functions

Monitor Jobs and **Save Results** as a New Table in BigQuery

- Track completed and ongoing jobs
- See the data quality metrics for transformed datasets
- View histograms with summary statistics for each field



Job Results

100%	0%	0%	247	307,483
Valid	Mismatched	Missing	Columns	Rows

Job Status	Complete	Launch Time	Today at 6:09 AM	CSV	347....
Job ID	5583	Finish Time	Today at 6:55 AM		
Environment	Google Dataflow	Duration	an hour		
Dataset	irs_990_2016				

	elf	EIN	tax_pd	calendar_year_filed	subseccd	s501c3c
Valid	307,483	307,483	307,483	307,483	307,483	
Mismatched	0	0	0	0	0	
Empty	0	0	0	0	0	

elf — Top 2 values

Y	222,375
N	85,108

	EIN	tax_pd	calendar_year_filed
Minimum	10,018,922	200,003	Mar 1 2000
Lower quartile	241,388,855	201,506	Jun 8 2015
Median	420,618,975	201,512	Dec 1 2015
Upper quartile	620,664,667	201,512	Dec 1 2015
Maximum	996,086,871	201,611	Nov 1 2016

subseccd — Top 20 values

03	229,549
06	20,570
04	11,561
05	10,047
07	10,043
09	5,397
08	4,707
19	4,511
12	2,886
02	2,068
13	1,958
14	1,712
10	1,545
25	593
15	187
17	67
29	39
23	11
26	10
16	7

s501c3c — Top 2 values

Y
N

Google Cloud

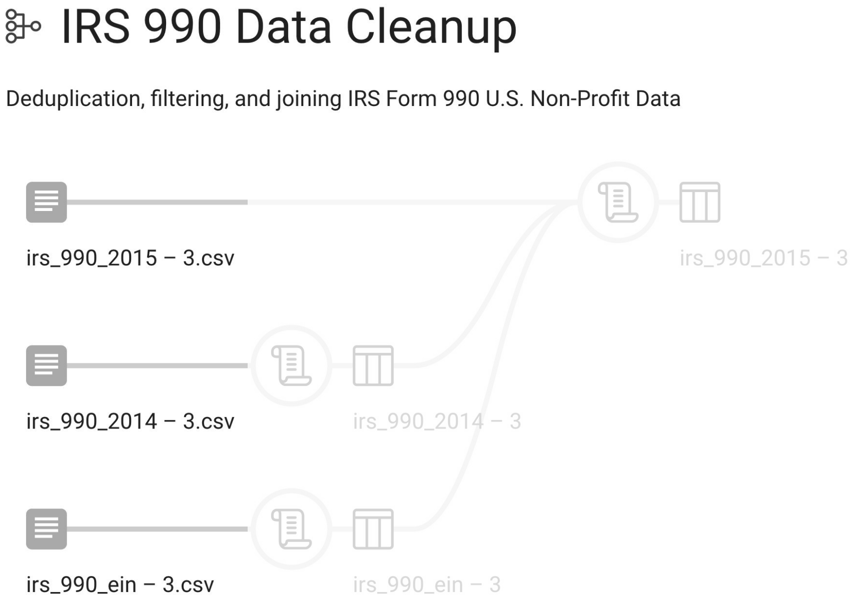
Lab 4a
Explore and Load Data with Cloud Dataprep

Google Cloud

Transform your data with Cloud Dataprep

Cloud Dataprep is Google's self-service data preparation tool.

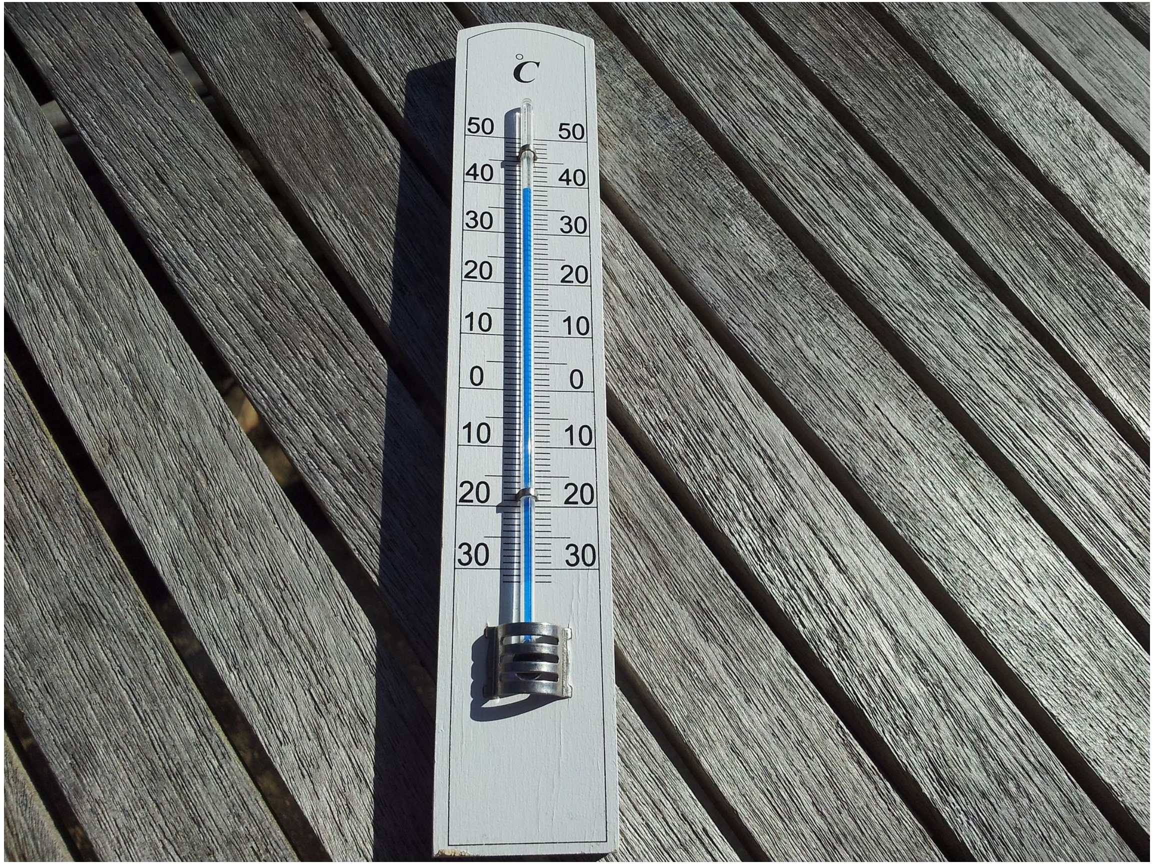
In the first part of this lab, we will load data sources as part of a new flow.



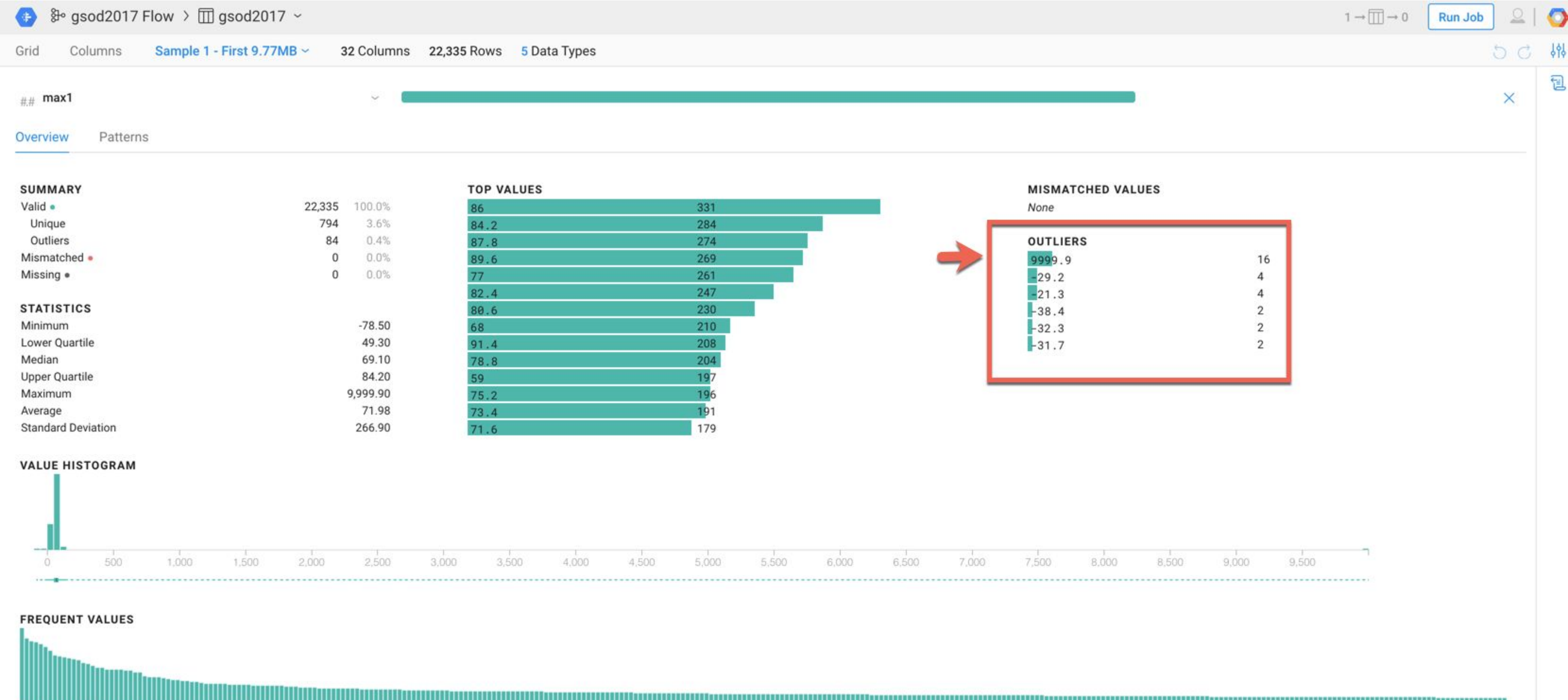
IRS 990 Data Cleanup

Deduplication, filtering, and joining IRS Form 990 U.S. Non-Profit Data

Cleaning NOAA Temperature Data with Cloud Dataprep



Using **Column Details** Statistics Reveals Outlier Max Temperature



SUMMARY

Valid ●	22,335	100.0%
Unique	794	3.6%
Outliers	84	0.4%
Mismatched ●	0	0.0%
Missing ●	0	0.0%

STATISTICS

Minimum	-78.50
Lower Quartile	49.30
Median	69.10
Upper Quartile	84.20
Maximum	9,999.90
Average	71.98
Standard Deviation	266.90

TOP VALUES

86	331
84.2	284
87.8	274
89.6	269
77	261
82.4	247
80.6	230
68	210
91.4	208
78.8	204
59	197
75.2	196
73.4	191
71.6	179

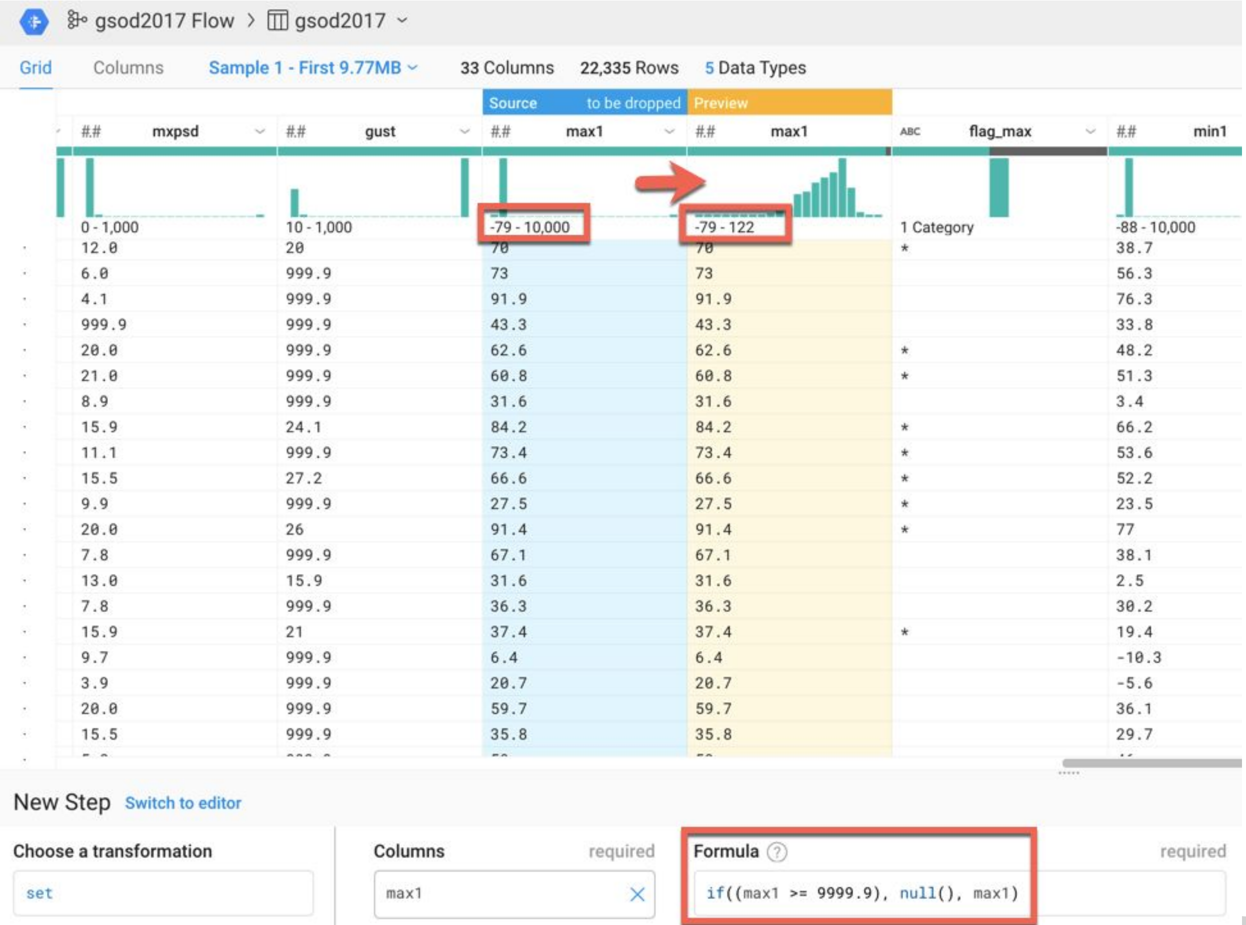
MISMATCHED VALUES

None

OUTLIERS

9999.9	16
-29.2	4
-21.3	4
-38.4	2
-32.3	2
-31.7	2

Set the Anomalous 9999.9 Temp Value to NULL with a Formula



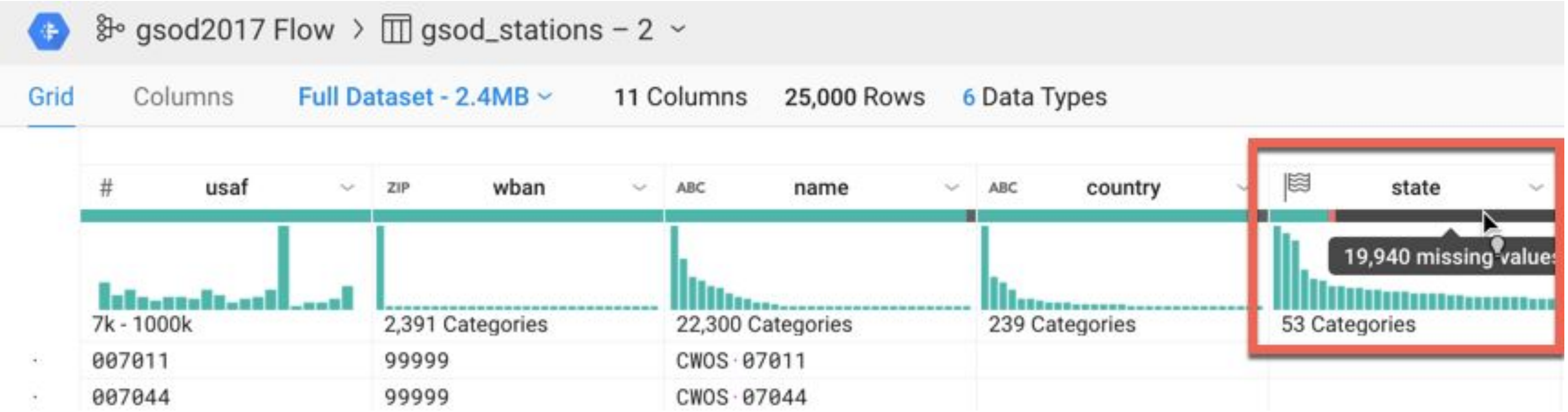
## mxpsd	## gust	## max1 (Source – to be dropped)	## max1 (Preview)	ABC flag_max	## min1
0 - 1,000	10 - 1,000	-79 - 10,000	-79 - 122	1 Category	-88 - 10,000
12.0	20	70	70	*	38.7
6.0	999.9	73	73		56.3
4.1	999.9	91.9	91.9		76.3
999.9	999.9	43.3	43.3		33.8
20.0	999.9	62.6	62.6	*	48.2
21.0	999.9	60.8	60.8	*	51.3
8.9	999.9	31.6	31.6		3.4
15.9	24.1	84.2	84.2	*	66.2
11.1	999.9	73.4	73.4	*	53.6
15.5	27.2	66.6	66.6	*	52.2
9.9	999.9	27.5	27.5	*	23.5
20.0	26	91.4	91.4	*	77
7.8	999.9	67.1	67.1		38.1
13.0	15.9	31.6	31.6		2.5
7.8	999.9	36.3	36.3		30.2
15.9	21	37.4	37.4	*	19.4
9.7	999.9	6.4	6.4		-10.3
3.9	999.9	20.7	20.7		-5.6
20.0	999.9	59.7	59.7		36.1
15.5	999.9	35.8	35.8		29.7

New Step Switch to editor

Choose a transformation	Columns (required)	Formula (required)
set	max1	`if((max1 >= 9999.9), null(), max1)`

Google Cloud

Looking at the **Data Quality Bar** shows many **States missing**



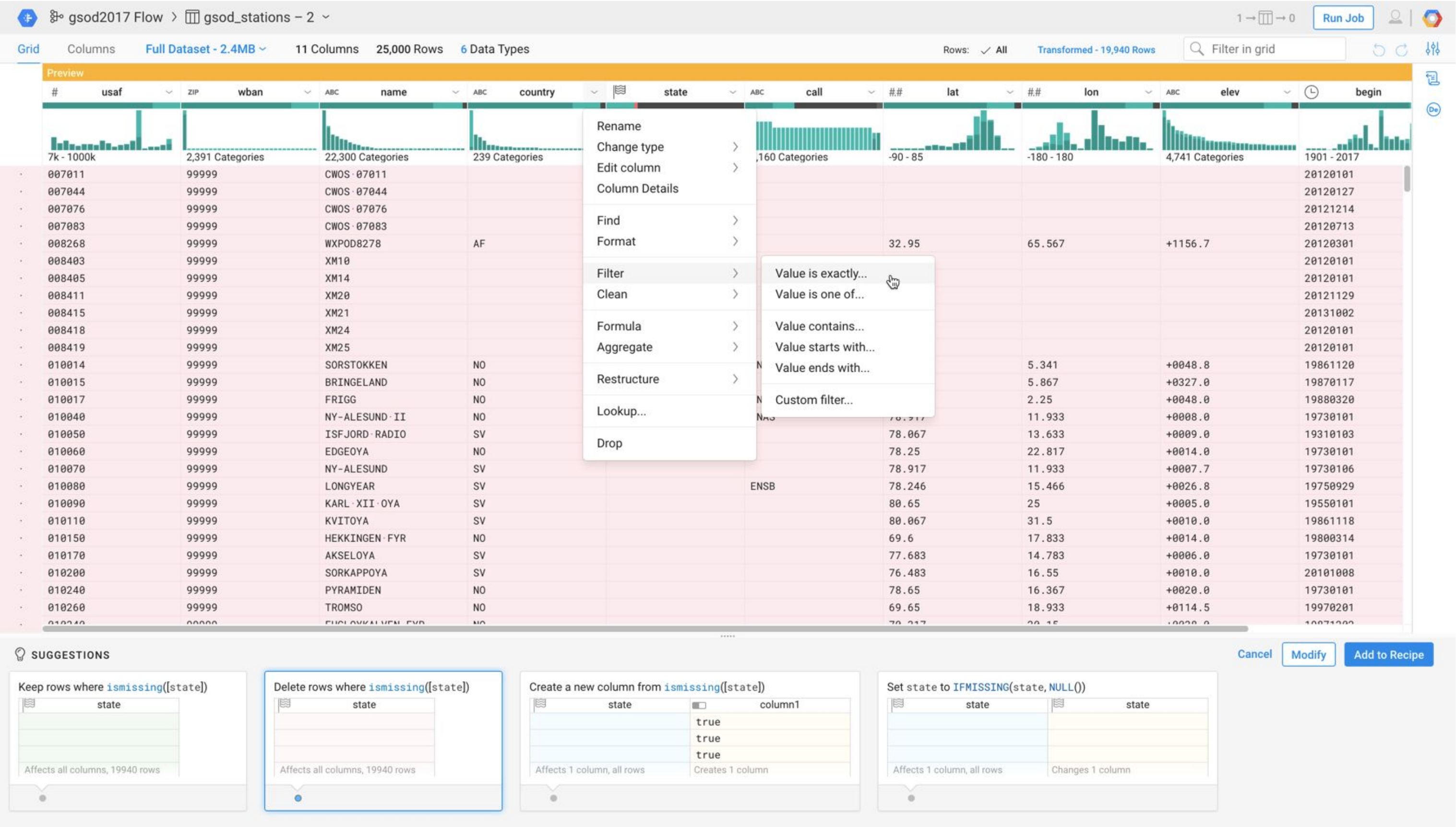
gsod2017 Flow > gsod_stations – 2

Grid | Columns | Full Dataset - 2.4MB | 11 Columns | 25,000 Rows | 6 Data Types

# usaf	ZIP wban	ABC name	ABC country	state
7k - 1000k	2,391 Categories	22,300 Categories	239 Categories	53 Categories
007011	99999	CWOS 07011		
007044	99999	CWOS 07044		

19,940 missing value

Filter on U.S. Only Weather Recordings



usaf	wban	name	country	state	call	lat	lon	elev	begin
007011	99999	CWOS 07011							20120101
007044	99999	CWOS 07044							20120127
007076	99999	CWOS 07076							20121214
007083	99999	CWOS 07083							20120713
008268	99999	WXPOD8278	AF			32.95	65.567	+1156.7	20120301
008403	99999	XM10							20120101
008405	99999	XM14							20120101
008411	99999	XM20							20121129
008415	99999	XM21							20131002
008418	99999	XM24							20120101
008419	99999	XM25							20120101
010014	99999	SORSTOKKEN	NO		N		5.341	+0048.8	19861120
010015	99999	BRINGELAND	NO				5.867	+0327.0	19870117
010017	99999	FRIGG	NO		N		2.25	+0048.0	19880320
010040	99999	NY-ALESUND II	NO		NAS	78.917	11.933	+0008.0	19730101
010050	99999	ISFJORD RADIO	SV			78.067	13.633	+0009.0	19310103
010060	99999	EDGEOYA	NO			78.25	22.817	+0014.0	19730101
010070	99999	NY-ALESUND	SV			78.917	11.933	+0007.7	19730106
010080	99999	LONGYEAR	SV		ENSB	78.246	15.466	+0026.8	19750929
010090	99999	KARL XII OYA	SV			80.65	25	+0005.0	19550101
010110	99999	KVITOYA	SV			80.067	31.5	+0010.0	19861118
010150	99999	HEKKINGEN FYR	NO			69.6	17.833	+0014.0	19800314
010170	99999	AKSELOYA	SV			77.683	14.783	+0006.0	19730101
010200	99999	SORKAPPOYA	SV			76.483	16.55	+0010.0	20101008
010240	99999	PYRAMIDEN	NO			78.65	16.367	+0020.0	19730101
010260	99999	TROMSO	NO			69.65	18.933	+0114.5	19970201

Keep rows where ismissing([state])

state

Affects all columns, 19940 rows

Delete rows where ismissing([state])

state

Affects all columns, 19940 rows

Create a new column from ismissing([state])

state	column1
	true
	true
	true

Affects 1 column, all rows
Creates 1 column

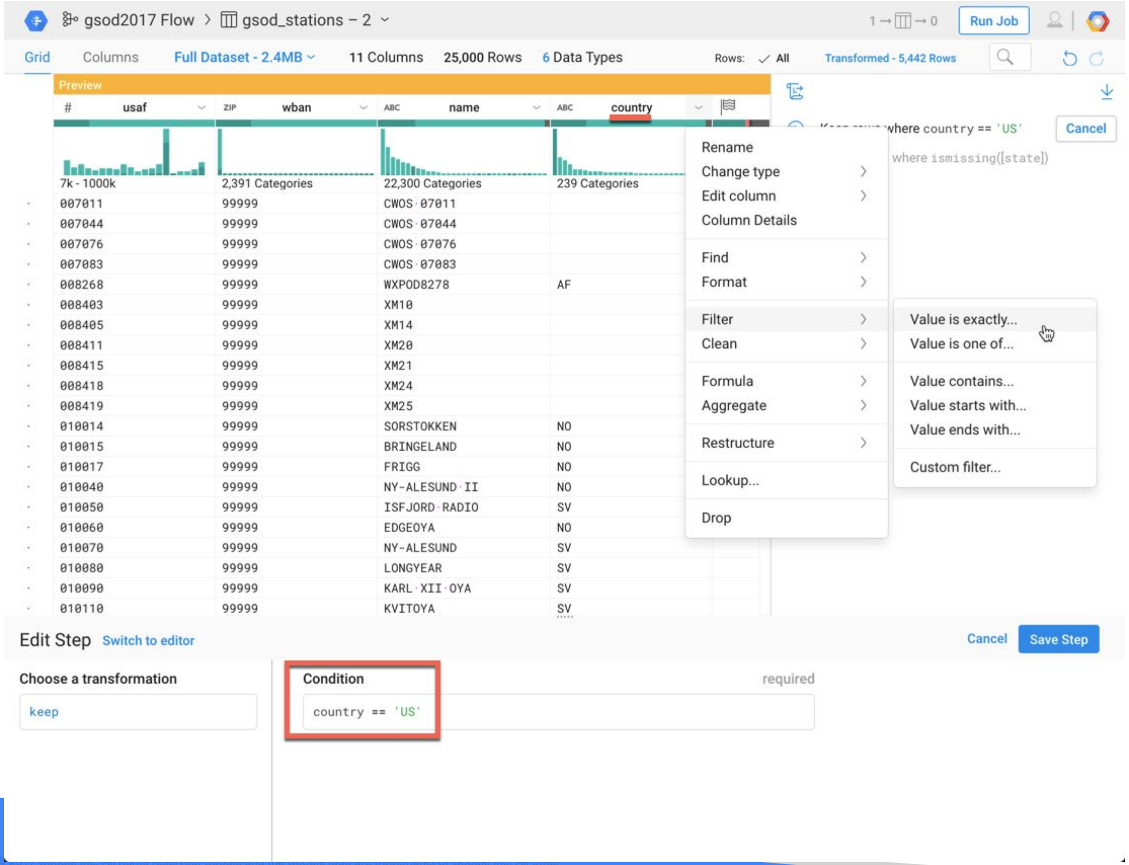
Set state to IFMISSING(state, NULL())

state	state

Affects 1 column, all rows
Changes 1 column

Google Cloud

Keep only U.S. Only Weather Recordings



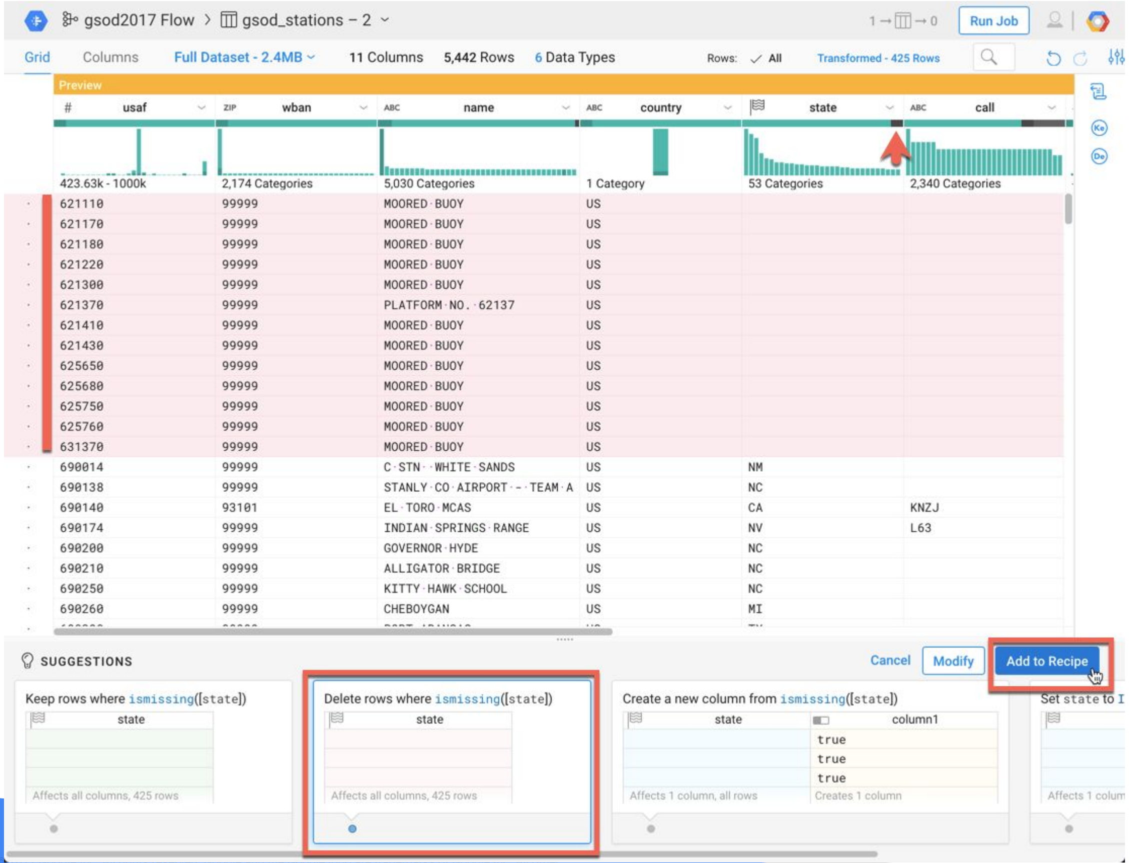
usaf	wban	name	country
7k - 1000k	2,391 Categories	22,300 Categories	239 Categories
007011	99999	CWOS 07011	
007044	99999	CWOS 07044	
007076	99999	CWOS 07076	
007083	99999	CWOS 07083	
008268	99999	WXPOD8278	AF
008403	99999	XM10	
008405	99999	XM14	
008411	99999	XM20	
008415	99999	XM21	
008418	99999	XM24	
008419	99999	XM25	
010014	99999	SORSTOKKEN	NO
010015	99999	BRINGELAND	NO
010017	99999	FRIGG	NO
010040	99999	NY-ALESUND II	NO
010050	99999	ISFJORD RADIO	SV
010060	99999	EDGEOYA	NO
010070	99999	NY-ALESUND	SV
010080	99999	LONGYEAR	SV
010090	99999	KARL XII OYA	SV
010110	99999	KVITOYA	SV

Edit Step Switch to editor

Choose a transformation: keep

Condition (required): `country == 'US'`

Delete Missing Data for the State Field



usaf	wban	name	country	state	call
423.63k - 1000k	2,174 Categories	5,030 Categories	1 Category	53 Categories	2,340 Categories
621110	99999	MOORED·BUOY	US		
621170	99999	MOORED·BUOY	US		
621180	99999	MOORED·BUOY	US		
621220	99999	MOORED·BUOY	US		
621300	99999	MOORED·BUOY	US		
621370	99999	PLATFORM·NO.·62137	US		
621410	99999	MOORED·BUOY	US		
621430	99999	MOORED·BUOY	US		
625650	99999	MOORED·BUOY	US		
625680	99999	MOORED·BUOY	US		
625750	99999	MOORED·BUOY	US		
625760	99999	MOORED·BUOY	US		
631370	99999	MOORED·BUOY	US		
690014	99999	C·STN··WHITE·SANDS	US	NM	
690138	99999	STANLY·CO·AIRPORT·-·TEAM·A	US	NC	
690140	93101	EL·TORO·MCAS	US	CA	KNZJ
690174	99999	INDIAN·SPRINGS·RANGE	US	NV	L63
690200	99999	GOVERNOR·HYDE	US	NC	
690210	99999	ALLIGATOR·BRIDGE	US	NC	
690250	99999	KITTY·HAWK·SCHOOL	US	NC	
690260	99999	CHEBOYGAN	US	MI	

Keep rows where ismissing([state])
state
Affects all columns, 425 rows

Delete rows where ismissing([state])
state
Affects all columns, 425 rows

Create a new column from ismissing([state])

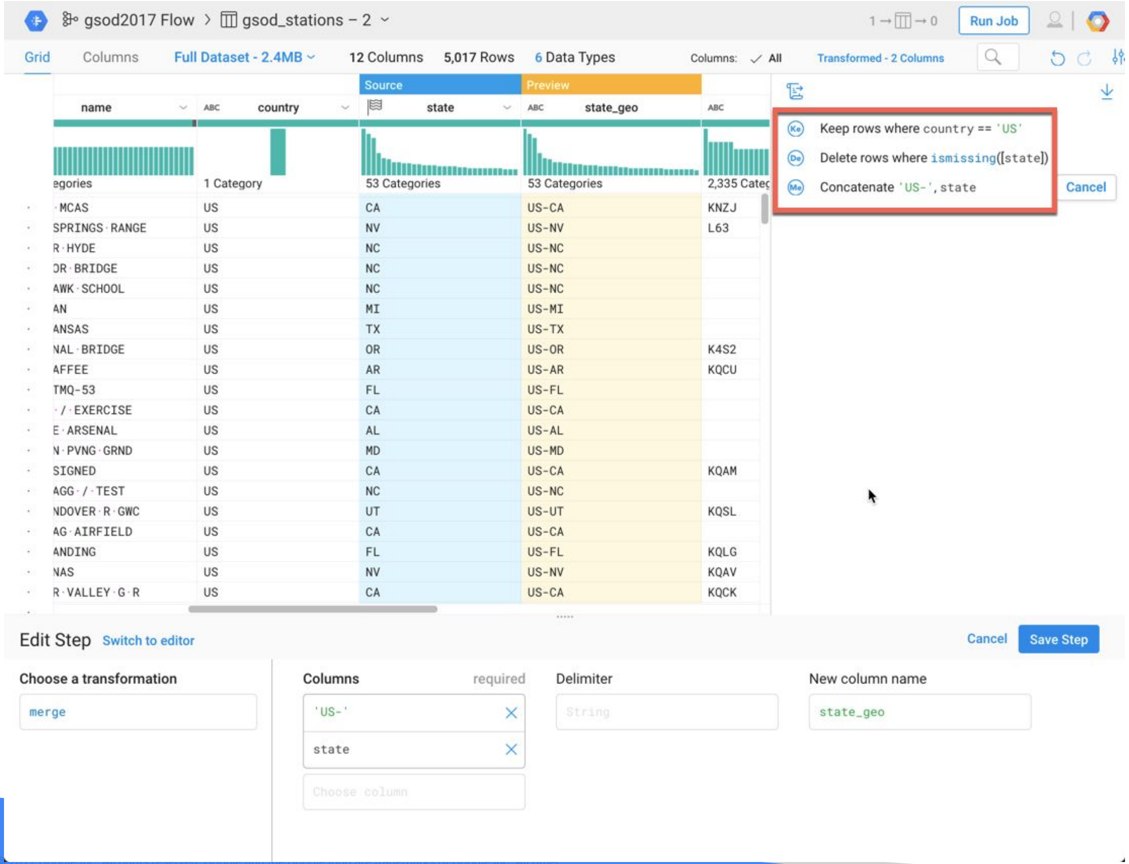
state	column1
	true
	true
	true

Affects 1 column, all rows
Creates 1 column

Set state to I
Affects 1 colum

- Browse through automatic **suggestion** cards for transformation
- **Modify** to customize your own logic
- **Add to Recipe** when ready

Review Final Recipe and Save



		Source	Preview	
name	ABC country	state	ABC state_geo	ABC
egories	1 Category	53 Categories	53 Categories	2,335 Categ
·MCAS	US	CA	US-CA	KNZJ
SPRINGS·RANGE	US	NV	US-NV	L63
R·HYDE	US	NC	US-NC	
OR·BRIDGE	US	NC	US-NC	
AWK·SCHOOL	US	NC	US-NC	
AN	US	MI	US-MI	
ANSAS	US	TX	US-TX	
NAL·BRIDGE	US	OR	US-OR	K4S2
AFFEE	US	AR	US-AR	KQCU
TMQ-53	US	FL	US-FL	
·/·EXERCISE	US	CA	US-CA	
E·ARSENAL	US	AL	US-AL	
N·PVNG·GRND	US	MD	US-MD	
SIGNED	US	CA	US-CA	KQAM
AGG·/·TEST	US	NC	US-NC	
NDOVER·R·GWC	US	UT	US-UT	KQSL
AG·AIRFIELD	US	CA	US-CA	
ANDING	US	FL	US-FL	KQLG
NAS	US	NV	US-NV	KQAV
R·VALLEY·G·R	US	CA	US-CA	KQCK

1. Keep rows where country == 'US'
2. Delete rows where ismissing([state])
3. Concatenate 'US-', state

Edit Step Switch to editor Cancel Save Step

Choose a transformation: merge

Columns (required): 'US-' ✕ / state ✕ / Choose column

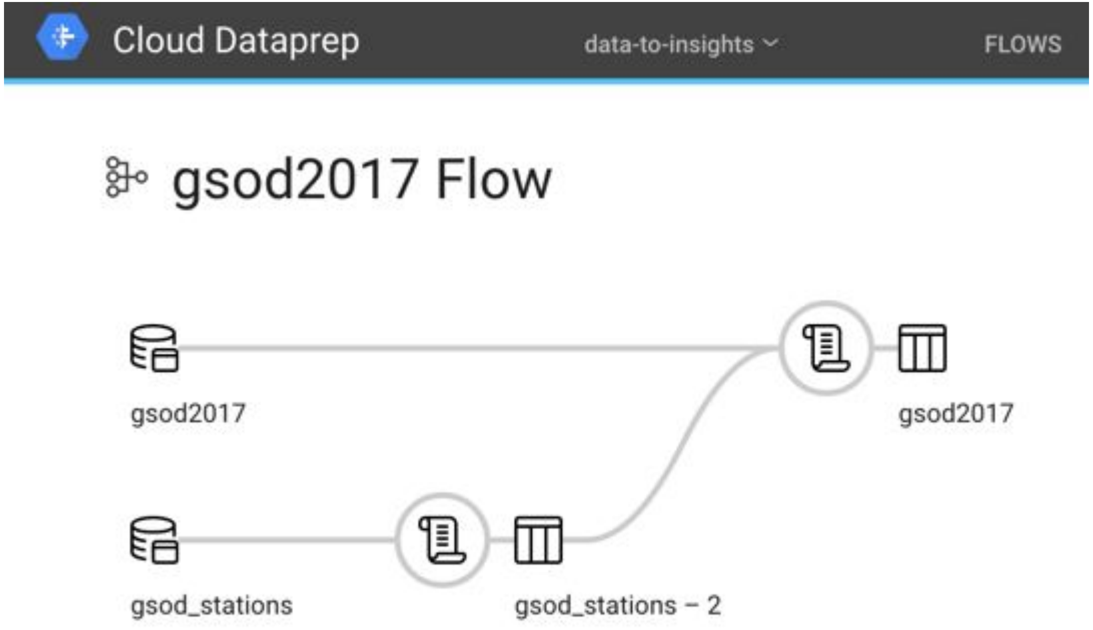
Delimiter: String

New column name: state_geo

- Toggle open the right side bar to view the steps in your recipe
- Modify or remove steps as needed
- Click **Run Job** when you want to Execute

Google Cloud

Run the Flow which includes our Recipes and Outputs a Table



Summary: Create clean datasets with SQL and/or Cloud Dataprep

Dataset integrity includes validity, accuracy, completeness, consistency, and uniformity

Explore your data to determine if there is heavy skew which could impact performance

Clean and transform your dataset by writing SQL statements

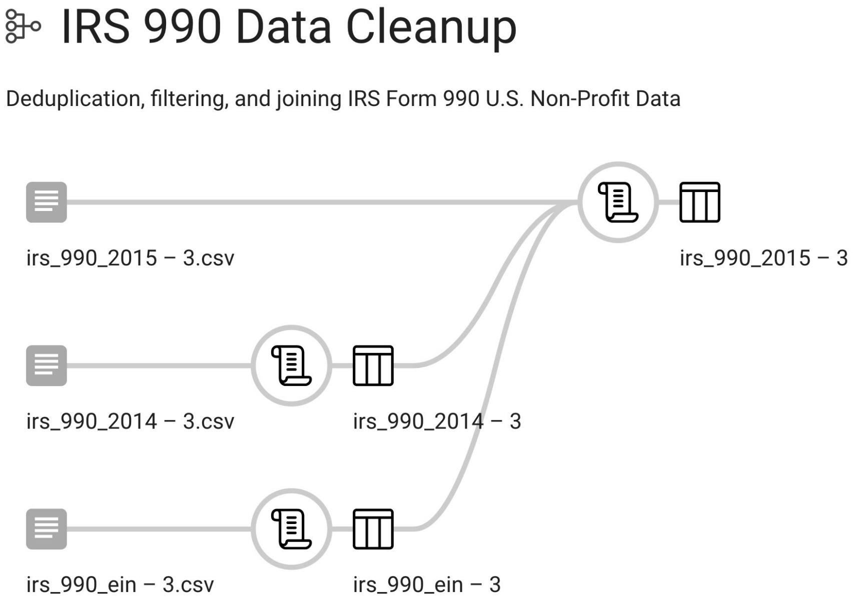
Clean and transform your dataset through the Cloud Dataprep UI

Lab 4b
Transform Data with Cloud Dataprep

Transform your data with Cloud Dataprep

In the second part of this lab, we will clean, merge, and join our IRS datasets together.

Afterward we will execute our first Cloud Dataprep pipeline job.



IRS 990 Data Cleanup

Deduplication, filtering, and joining IRS Form 990 U.S. Non-Profit Data