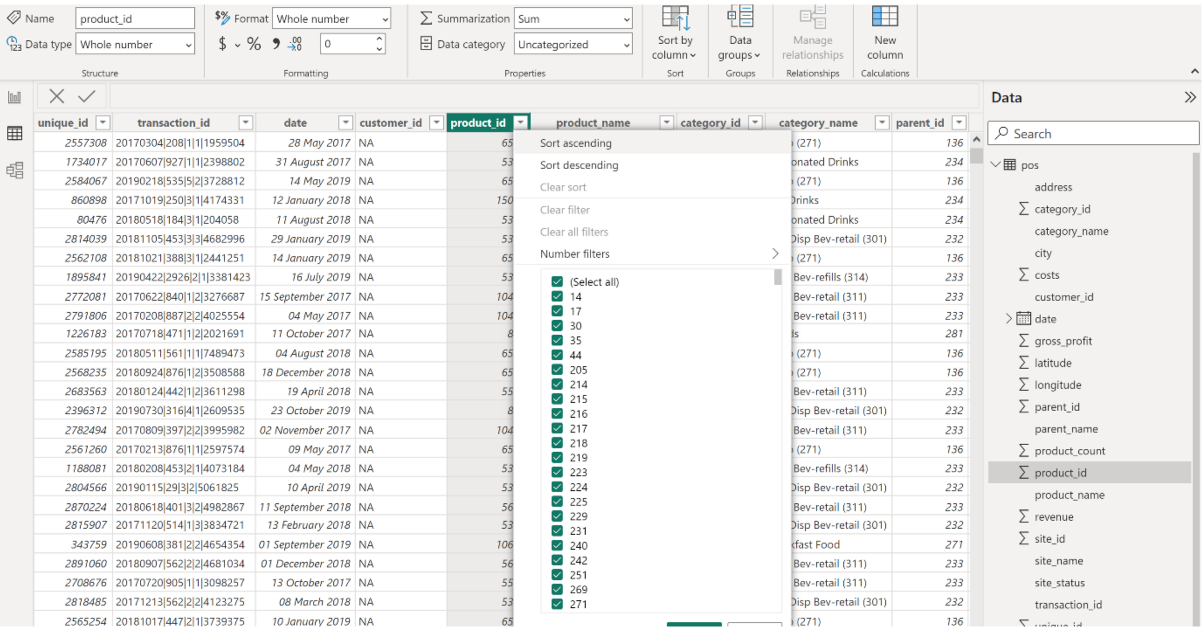
Power BI – Extract, Transform and Load

ETL 1 – Examine Data with Power Query Editor

Load the data -> pos for (Point of Sale). The data is from a fictitious convenience store/gas station.

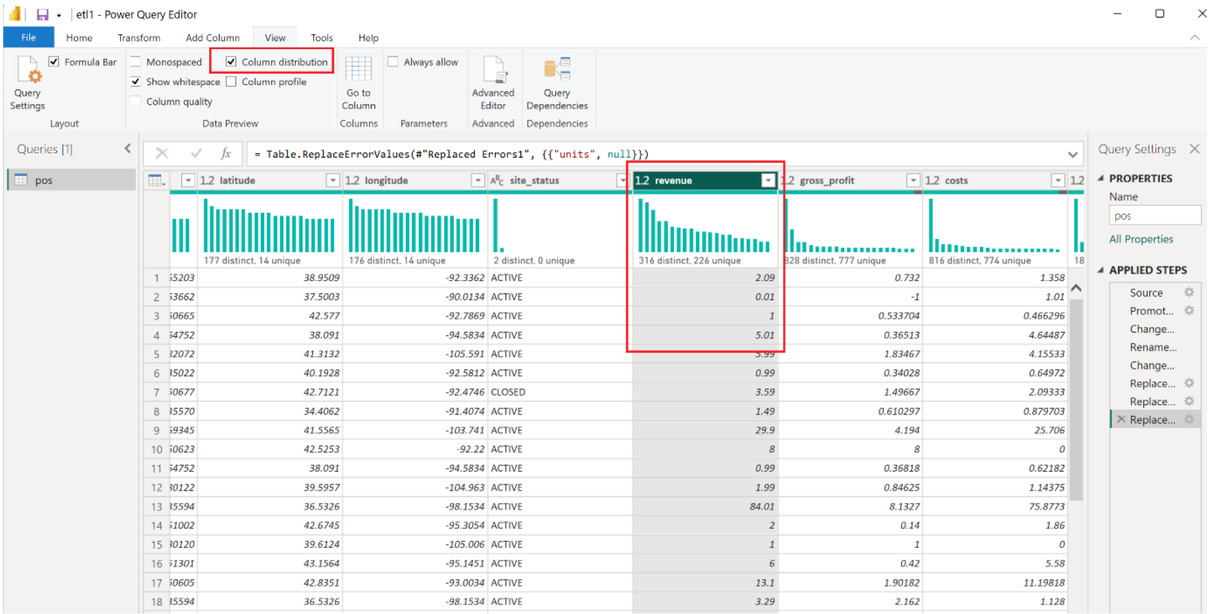


Name: product_id
Data type: Whole number
Format: Whole number
0
Summarization: Sum
Data category: Uncategorized

☑ (Select all)
☑ 14
☑ 17
☑ 30
☑ 35
☑ 44
☑ 205
☑ 214
☑ 215
☑ 216
☑ 217
☑ 218
☑ 219
☑ 223
☑ 224
☑ 225
☑ 229
☑ 231
☑ 240
☑ 242
☑ 251
☑ 269
☑ 271

unique_id	transaction_id	date	customer_id	product_id	product_name	category_id	category_name	parent_id
2557308	20170304|208|1|1|1959504	28 May 2017	NA	65			(271)	136
1734017	20170607|927|1|1|2398802	31 August 2017	NA	53			onated Drinks	234
2584067	20190218|535|5|2|3728812	14 May 2019	NA	65			(271)	136
860898	20171019|250|3|1|4174331	12 January 2018	NA	150			Drinks	234
80476	20180518|184|3|1|204058	11 August 2018	NA	53			onated Drinks	234
2814039	20181105|453|3|3|4682996	29 January 2019	NA	53			Disp Bev-retail (301)	232
2562108	20181021|388|3|1|2441251	14 January 2019	NA	65			(271)	136
1895841	20190422|2926|2|1|3381423	16 July 2019	NA	53			Bev-refills (314)	233
2772081	20170622|840|1|2|3276687	15 September 2017	NA	104			Bev-retail (311)	233
2791806	20170208|887|2|2|4025554	04 May 2017	NA	104			Bev-retail (311)	233
1226183	20170718|471|1|2|2021691	11 October 2017	NA	8			ls	281
2585195	20180511|561|1|1|7489473	04 August 2018	NA	65			(271)	136
2568235	20180924|876|1|2|3508588	18 December 2018	NA	65			(271)	136
2683563	20180124|442|1|2|3611298	19 April 2018	NA	55			Bev-retail (311)	233
2396312	20190730|316|4|1|2609535	23 October 2019	NA	8			Disp Bev-retail (301)	232
2782494	20170809|397|2|2|3995982	02 November 2017	NA	104			Bev-retail (311)	233
2561260	20170213|876|1|1|2597574	09 May 2017	NA	65			(271)	136
1188081	20180208|453|2|1|4073184	04 May 2018	NA	53			Bev-refills (314)	233
2804566	20190115|29|3|2|5061825	10 April 2019	NA	53			Disp Bev-retail (301)	232
2870224	20180618|401|3|2|4982867	11 September 2018	NA	56			Bev-retail (311)	233
2815907	20171120|514|1|3|3834721	13 February 2018	NA	53			Disp Bev-retail (301)	232
343759	20190608|381|2|2|4654354	01 September 2019	NA	106			kfast Food	271
2891060	20180907|562|2|2|4681034	01 December 2018	NA	56			Bev-retail (311)	233
2708676	20170720|905|1|1|3098257	13 October 2017	NA	55			Bev-retail (311)	233
2818485	20171213|562|2|2|4123275	08 March 2018	NA	53			Disp Bev-retail (301)	232
2565254	20181017|447|2|1|3739375	10 January 2019	NA	65			(271)	136

View a column distribution by clicking on a particular column, clicking on ‘View’ and selecting ‘Column distribution’. Transformation steps are shown on right-hand side under ‘Applied Steps’.



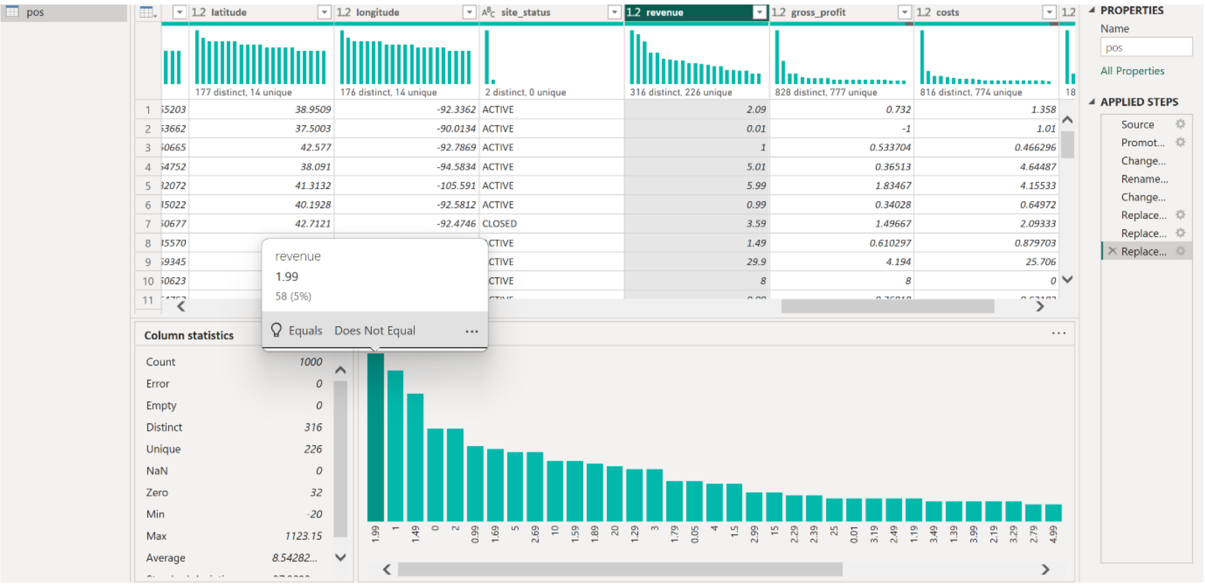
☑ Formula Bar
☐ Monospaced
☑ Column distribution
☑ Show whitespace
☐ Column profile
☐ Column quality
☐ Always allow

Name: pos

```
= Table.ReplaceErrorValues(#"Replaced Errors1", {{"units", null}})
```

		1.2 latitude	1.2 longitude	A^BC site_status	1.2 revenue	1.2 gross_profit	1.2 costs
		177 distinct, 14 unique	176 distinct, 14 unique	2 distinct, 0 unique	316 distinct, 226 unique	828 distinct, 777 unique	816 distinct, 774 unique
1	55203	38.9509	-92.3362	ACTIVE	2.09	0.732	1.358
2	53662	37.5003	-90.0134	ACTIVE	0.01	-1	1.01
3	50665	42.577	-92.7869	ACTIVE	1	0.533704	0.466296
4	54752	38.091	-94.5834	ACTIVE	5.01	0.36513	4.64487
5	32072	41.3132	-105.591	ACTIVE	5.99	1.83467	4.15533
6	35022	40.1928	-92.5812	ACTIVE	0.99	0.34028	0.64972
7	50677	42.7121	-92.4746	CLOSED	3.59	1.49667	2.09333
8	35570	34.4062	-91.4074	ACTIVE	1.49	0.610297	0.879703
9	59345	41.5565	-103.741	ACTIVE	29.9	4.194	25.706
10	50623	42.5253	-92.22	ACTIVE	8	8	0
11	54752	38.091	-94.5834	ACTIVE	0.99	0.36818	0.62182
12	30122	39.5957	-104.963	ACTIVE	1.99	0.84625	1.14375
13	35594	36.5326	-98.1534	ACTIVE	84.01	8.1327	75.8773
14	51002	42.6745	-95.3054	ACTIVE	2	0.14	1.86
15	30120	39.6124	-105.006	ACTIVE	1	1	0
16	51301	43.1564	-95.1451	ACTIVE	6	0.42	5.58
17	50605	42.8351	-93.0034	ACTIVE	13.1	1.90182	11.19818
18	35594	36.5326	-98.1534	ACTIVE	3.29	2.162	1.128

Clicking on ‘View’, then checking ‘Column profile’ gives the following view

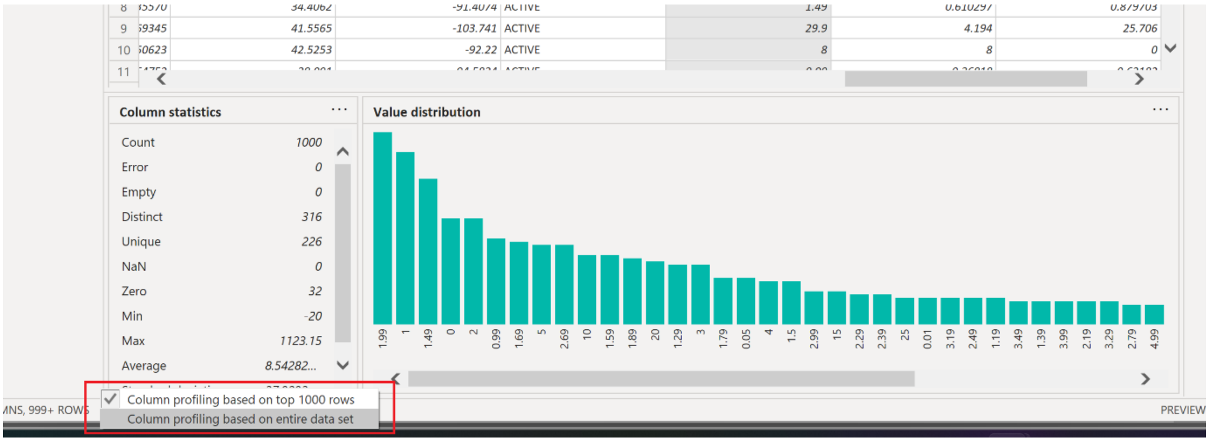


		1.2 latitude	1.2 longitude	A^B_C site_status	1.2 revenue	1.2 gross_profit	1.2 costs
		177 distinct, 14 unique	176 distinct, 14 unique	2 distinct, 0 unique	316 distinct, 226 unique	828 distinct, 777 unique	816 distinct, 774 unique
1	55203	38.9509	-92.3362	ACTIVE	2.09	0.732	1.358
2	53662	37.5003	-90.0134	ACTIVE	0.01	-1	1.01
3	50665	42.577	-92.7869	ACTIVE	1	0.533704	0.466296
4	54752	38.091	-94.5834	ACTIVE	5.01	0.36513	4.64487
5	32072	41.3132	-105.591	ACTIVE	5.99	1.83467	4.15533
6	35022	40.1928	-92.5812	ACTIVE	0.99	0.34028	0.64972
7	50677	42.7121	-92.4746	CLOSED	3.59	1.49667	2.09333
8	35570			CTIVE	1.49	0.610297	0.879703
9	59345			CTIVE	29.9	4.194	25.706
10	50623			CTIVE	8	8	0

Column statistics

Count	1000
Error	0
Empty	0
Distinct	316
Unique	226
NaN	0
Zero	32
Min	-20
Max	1123.15
Average	8.54282...

We can base the column profiling on the entire dataset



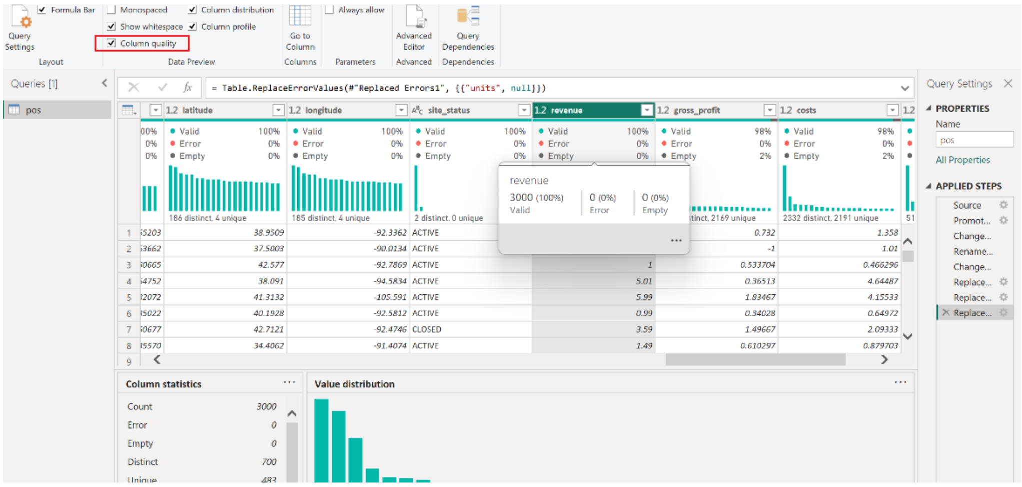
8	35570	34.4062	-91.4074	ACTIVE	1.49	0.610297	0.879703
9	59345	41.5565	-103.741	ACTIVE	29.9	4.194	25.706
10	50623	42.5253	-92.22	ACTIVE	8	8	0

Column statistics

Count	1000
Error	0
Empty	0
Distinct	316
Unique	226
NaN	0
Zero	32
Min	-20
Max	1123.15
Average	8.54282...

✓ Column profiling based on top 1000 rows
Column profiling based on entire data set

Now tick the ‘Column quality’ checkbox. This gives the percentage of Valid, Error and Empty entries



☑ Formula Bar
☐ Monospaced
☑ Show whitespace
☑ Column quality
☑ Column distribution
☑ Column profile
☐ Always allow

```
= Table.ReplaceErrorValues(#"Replaced Errors1", {{"units", null}})
```

		1.2 latitude	1.2 longitude	A^B_C site_status	1.2 revenue	1.2 gross_profit	1.2 costs
	00%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 98%	● Valid 98%
	0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%
	0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 2%	● Empty 2%
		186 distinct, 4 unique	185 distinct, 4 unique	2 distinct, 0 unique		tinct, 2169 unique	2332 distinct, 2191 unique
1	55203	38.9509	-92.3362	ACTIVE		0.732	1.358
2	53662	37.5003	-90.0134	ACTIVE		-1	1.01
3	50665	42.577	-92.7869	ACTIVE	1	0.533704	0.466296
4	54752	38.091	-94.5834	ACTIVE	5.01	0.36513	4.64487
5	32072	41.3132	-105.591	ACTIVE	5.99	1.83467	4.15533
6	35022	40.1928	-92.5812	ACTIVE	0.99	0.34028	0.64972
7	50677	42.7121	-92.4746	CLOSED	3.59	1.49667	2.09333
8	35570	34.4062	-91.4074	ACTIVE	1.49	0.610297	0.879703

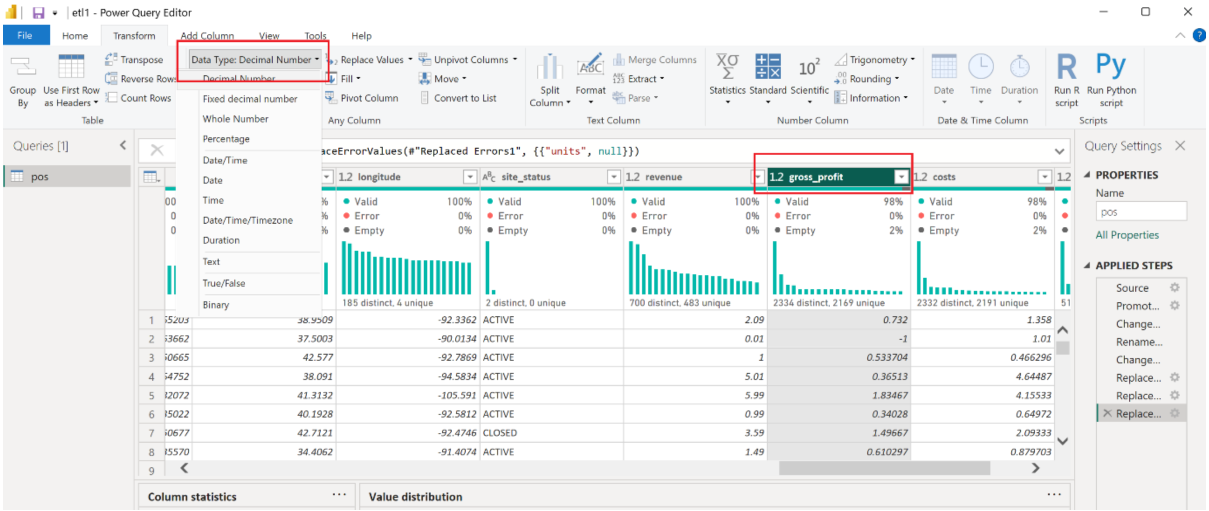
revenue

3000 (100%)	0 (0%)	0 (0%)
Valid	Error	Empty

Column statistics

Count	3000
Error	0
Empty	0
Distinct	700
Unique	483

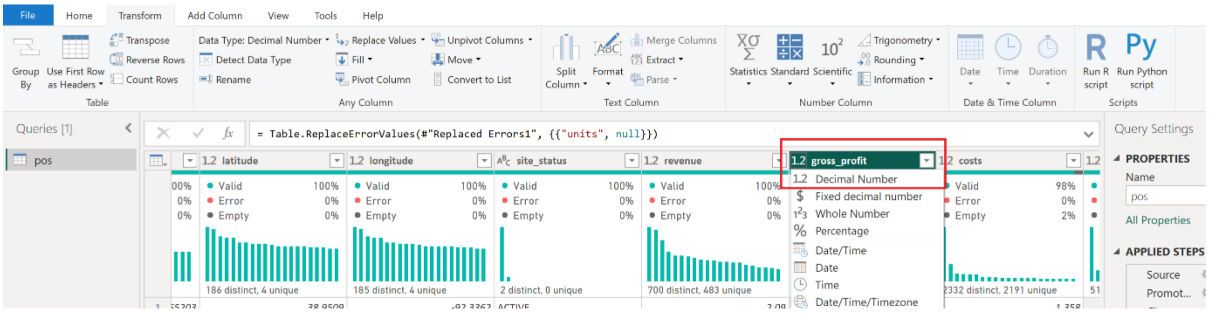
Change number columns to correct type using Data Type drop-down.



```
aceErrorValues(#"Replaced Errors1", {{"units", null}})
```

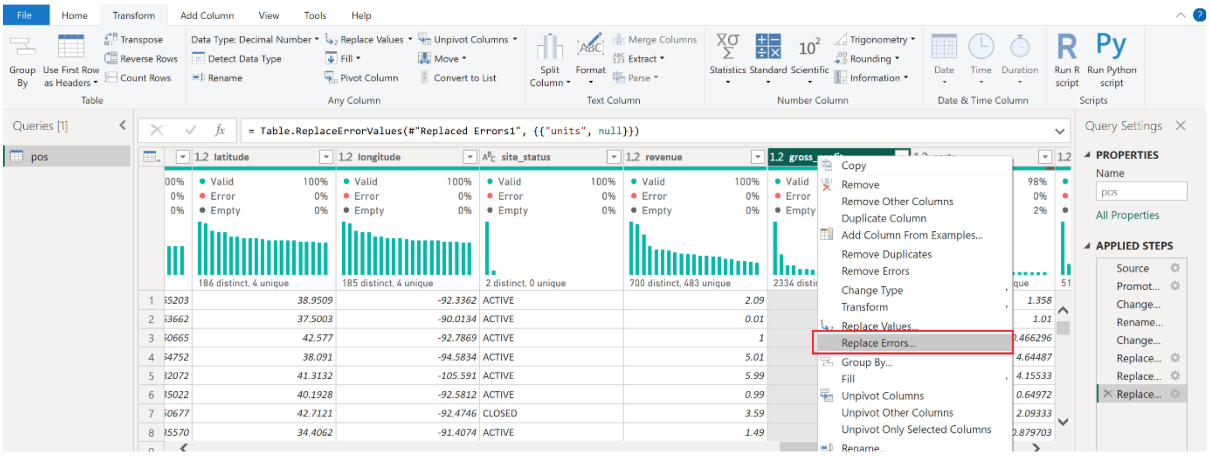
		1.2 longitude	site_status	1.2 revenue	1.2 gross_profit	1.2 costs
		Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 98% Error 0% Empty 2%	Valid 98% Error 0% Empty 2%
		185 distinct, 4 unique	2 distinct, 0 unique	700 distinct, 483 unique	2334 distinct, 2169 unique	2332 distinct, 2191 unique
1	38.9509	-92.3362	ACTIVE	2.09	0.732	1.358
2	37.5003	-90.0134	ACTIVE	0.01	-1	1.01
3	42.577	-92.7869	ACTIVE	1	0.533704	0.466296
4	38.091	-94.5834	ACTIVE	5.01	0.36513	4.64487
5	41.3132	-105.591	ACTIVE	5.99	1.83467	4.15533
6	40.1928	-92.5812	ACTIVE	0.99	0.34028	0.64972
7	42.7121	-92.4746	CLOSED	3.59	1.49667	2.09333
8	34.4062	-91.4074	ACTIVE	1.49	0.610297	0.879703

This can also be changed using the type symbol on the column header



```
= Table.ReplaceErrorValues(#"Replaced Errors1", {{"units", null}})
```

Replace errors with null

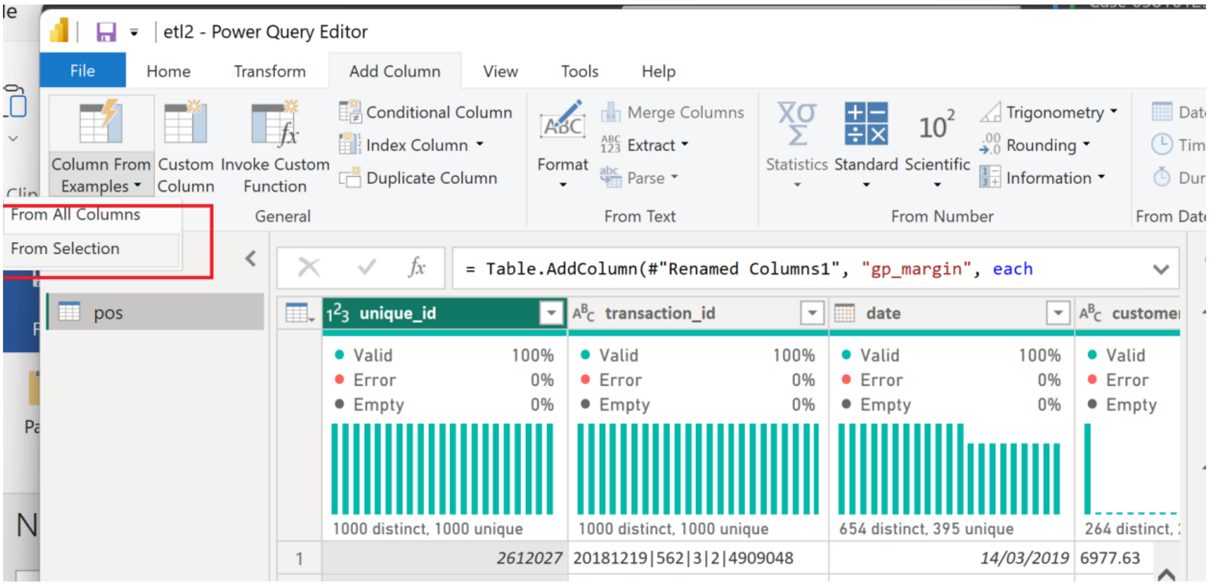


```
= Table.ReplaceErrorValues(#"Replaced Errors1", {{"units", null}})
```

	1.2 latitude	1.2 longitude	site_status	1.2 revenue
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	186 distinct, 4 unique	185 distinct, 4 unique	2 distinct, 0 unique	700 distinct, 483 unique
1	38.9509	-92.3362	ACTIVE	2.09
2	37.5003	-90.0134	ACTIVE	0.01
3	42.577	-92.7869	ACTIVE	1
4	38.091	-94.5834	ACTIVE	5.01
5	41.3132	-105.591	ACTIVE	5.99
6	40.1928	-92.5812	ACTIVE	0.99
7	42.7121	-92.4746	CLOSED	3.59
8	34.4062	-91.4074	ACTIVE	1.49

ETL 2 – Dates and Calculated Columns

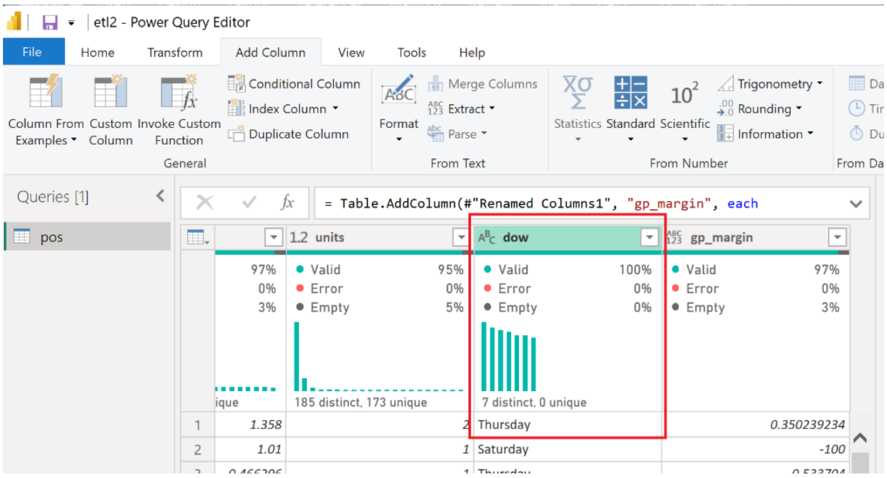
Add new column from selection



```
= Table.AddColumn(#"Renamed Columns1", "gp_margin", each
```

	unique_id	transaction_id	date	custome
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid Error Empty
	1000 distinct, 1000 unique	1000 distinct, 1000 unique	654 distinct, 395 unique	264 distinct,
1	2612027	20181219|562|3|2|4909048	14/03/2019	6977.63

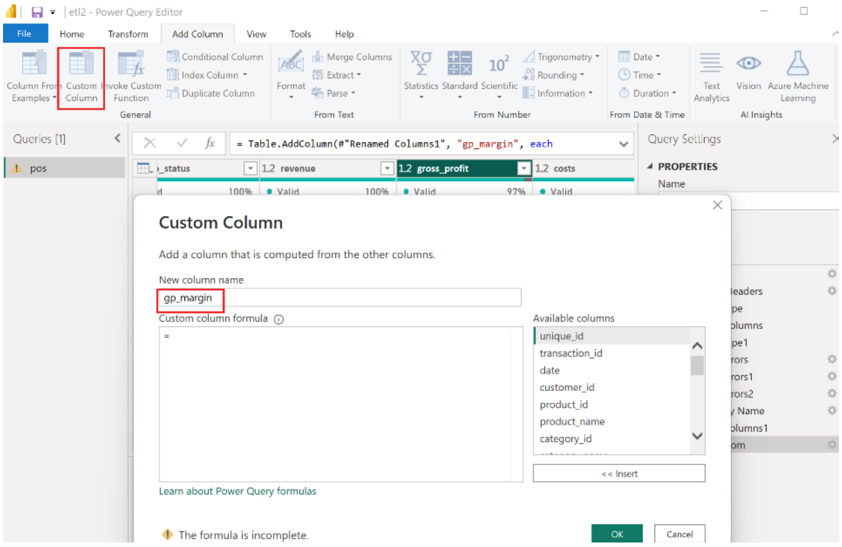
Add new column for Day of Week



```
= Table.AddColumn(#"Renamed Columns1", "gp_margin", each
```

		units	dow	gp_margin
	97% 0% 3%	Valid 95% Error 0% Empty 5%	Valid 100% Error 0% Empty 0%	Valid 97% Error 0% Empty 3%
	ique	185 distinct, 173 unique	7 distinct, 0 unique	
1	1.358	2	Thursday	0.350239234
2	1.01	1	Saturday	-100

Add new Custom Column for gross profit margin



```
= Table.AddColumn(#"Renamed Columns1", "gp_margin", each
```

Custom Column

Add a column that is computed from the other columns.

New column name: gp_margin

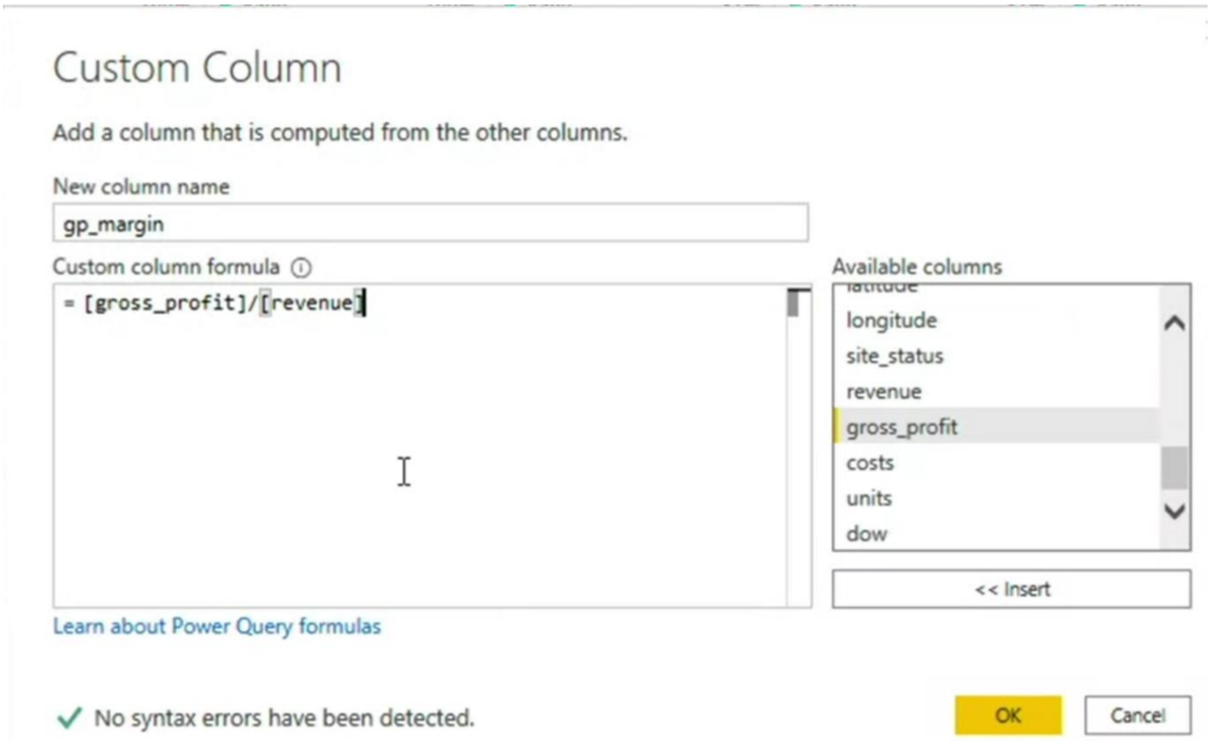
Custom column formula: =

Available columns

Learn about Power Query formulas

The formula is incomplete.

Use custom formula to calculate value for column



Custom Column

Add a column that is computed from the other columns.

New column name: gp_margin

Custom column formula:

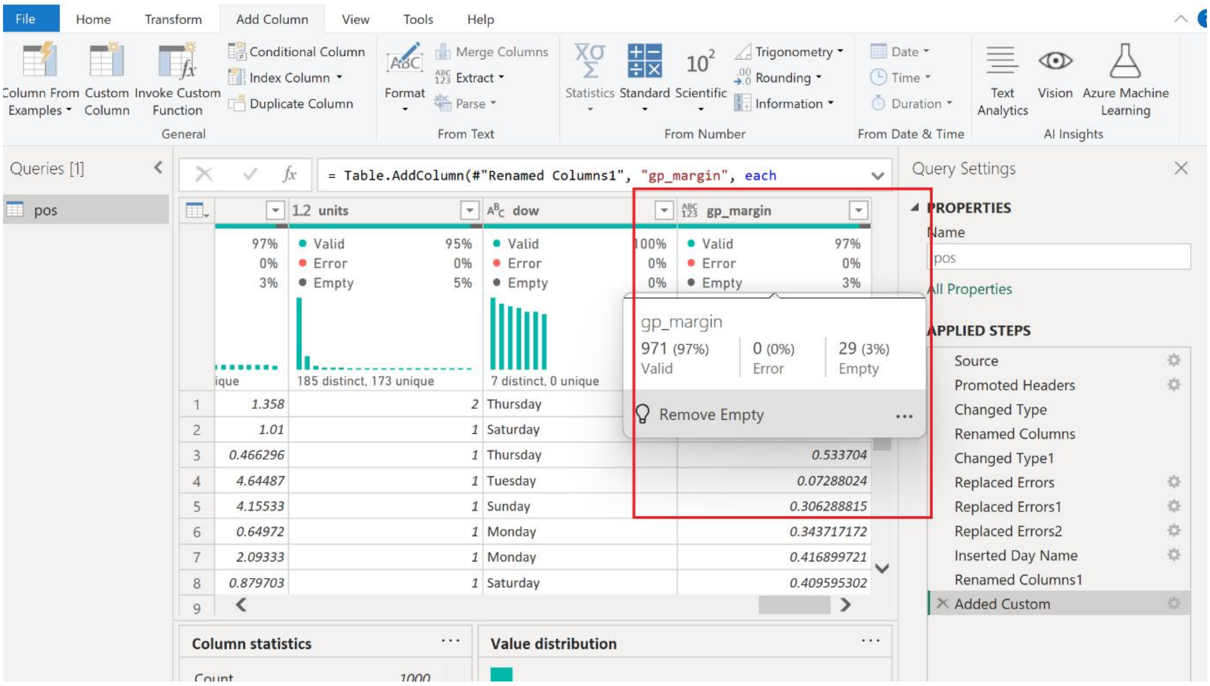
```
= [gross_profit]/[revenue]
```

Available columns

Learn about Power Query formulas

No syntax errors have been detected.

New column added



```
= Table.AddColumn(#"Renamed Columns1", "gp_margin", each
```

		1.2 units	A^BC dow	ABC 123 gp_margin
	● Valid 97% ● Error 0% ● Empty 3%	● Valid 95% ● Error 0% ● Empty 5%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 97% ● Error 0% ● Empty 3%
	ique	185 distinct, 173 unique	7 distinct, 0 unique	
1	1.358	2	Thursday	
2	1.01	1	Saturday	
3	0.466296	1	Thursday	0.533704
4	4.64487	1	Tuesday	0.07288024
5	4.15533	1	Sunday	0.306288815
6	0.64972	1	Monday	0.343717172
7	2.09333	1	Monday	0.416899721
8	0.879703	1	Saturday	0.409595302
9				

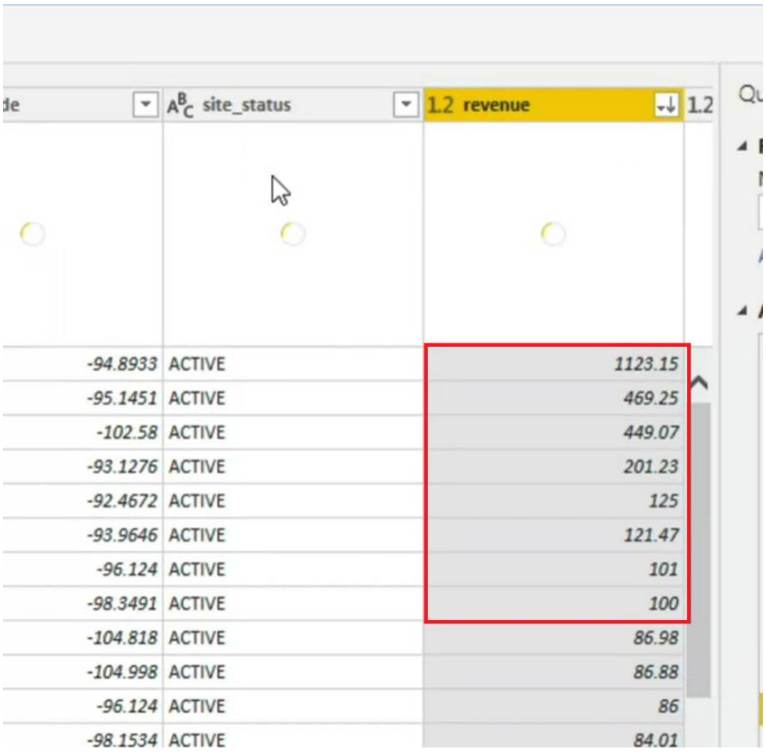
gp_margin
971 (97%) Valid | 0 (0%) Error | 29 (3%) Empty
Remove Empty

PROPERTIES
Name
pos
All Properties

APPLIED STEPS
- Source
- Promoted Headers
- Changed Type
- Renamed Columns
- Changed Type1
- Replaced Errors
- Replaced Errors1
- Replaced Errors2
- Inserted Day Name
- Renamed Columns1
- Added Custom

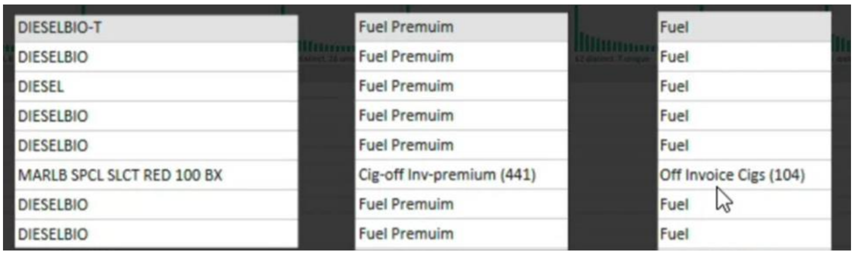
ETL 3 – Checking for and Eliminating Outliers

The highlighted rows are 3 standard deviations more than the mean (>87), therefore outliers.



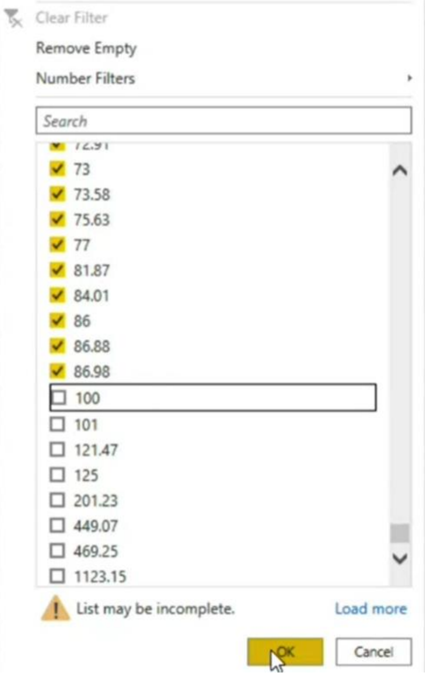
de	A^BC site_status	1.2 revenue
-94.8933	ACTIVE	1123.15
-95.1451	ACTIVE	469.25
-102.58	ACTIVE	449.07
-93.1276	ACTIVE	201.23
-92.4672	ACTIVE	125
-93.9646	ACTIVE	121.47
-96.124	ACTIVE	101
-98.3491	ACTIVE	100
-104.818	ACTIVE	86.98
-104.998	ACTIVE	86.88
-96.124	ACTIVE	86
-98.1534	ACTIVE	84.01

These outliers are related to fuel



DIESELBIO-T	Fuel Premuim	Fuel
DIESELBIO	Fuel Premuim	Fuel
DIESEL	Fuel Premuim	Fuel
DIESELBIO	Fuel Premuim	Fuel
DIESELBIO	Fuel Premuim	Fuel
MARLB SPCL SLCT RED 100 BX	Cig-off Inv-premium (441)	Off Invoice Cigs (104)
DIESELBIO	Fuel Premuim	Fuel
DIESELBIO	Fuel Premuim	Fuel

We can remove these outliers by manually deselecting them from the column



Clear Filter
Remove Empty
Number Filters

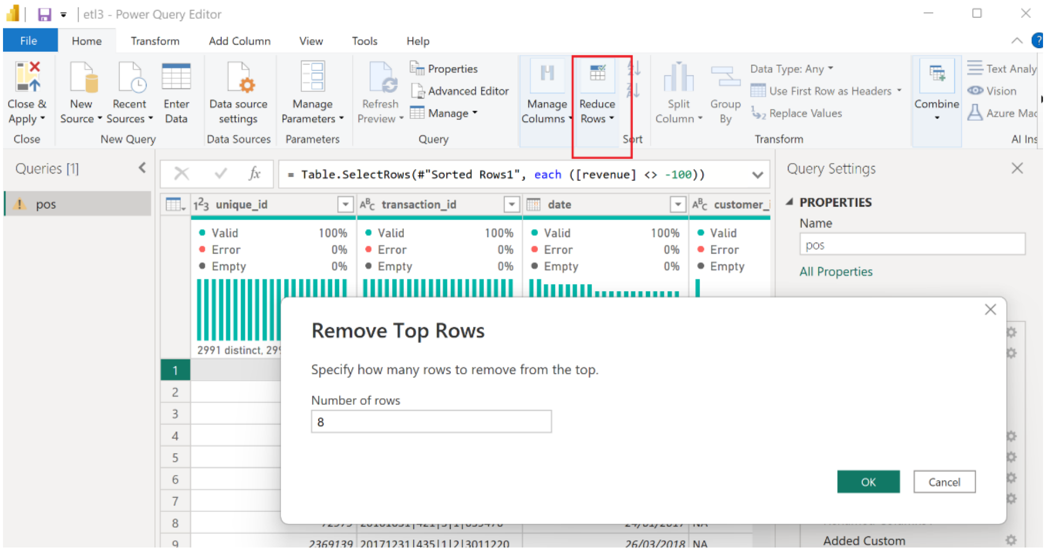
Search

- ☑ 72.91
- ☑ 73
- ☑ 73.58
- ☑ 75.63
- ☑ 77
- ☑ 81.87
- ☑ 84.01
- ☑ 86
- ☑ 86.88
- ☑ 86.98
- □ 100
- □ 101
- □ 121.47
- □ 125
- □ 201.23
- □ 449.07
- □ 469.25
- □ 1123.15

List may be incomplete. Load more

OK Cancel

Another way to do this is to go to ‘Reduce Rows’ in the Home tool and select ‘Remove Top Rows’ and specify the number of rows to remove. Note this will only work if you’ve sorted the data correctly.



```
= Table.SelectRows(#"Sorted Rows1", each ([revenue] <> -100))
```

unique_id	transaction_id	date	customer_
Valid 100%	Valid 100%	Valid 100%	Valid
Error 0%	Error 0%	Error 0%	Error
Empty 0%	Empty 0%	Empty 0%	Empty

Name: pos

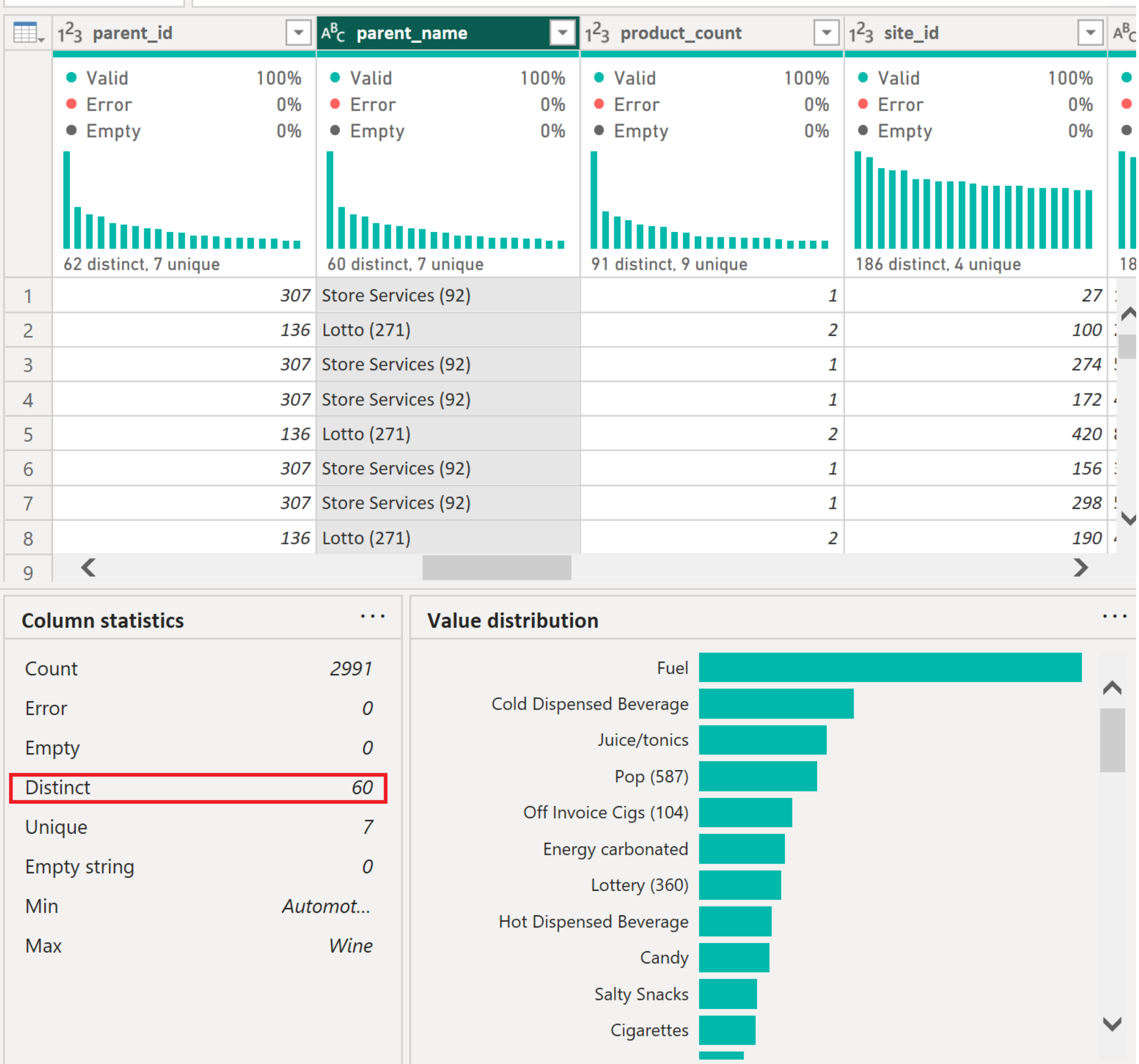
Remove Top Rows

Specify how many rows to remove from the top.

Number of rows: 8

OK Cancel

The column parent_name has high cardinality, meaning that there are 60 distinct types.



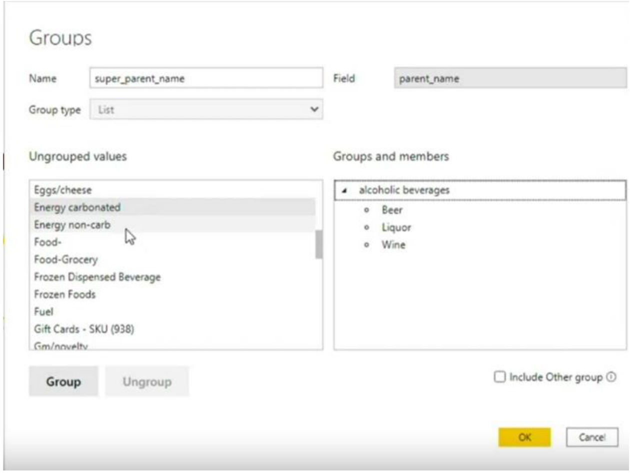
	parent_id	parent_name	product_count	site_id
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	62 distinct, 7 unique	60 distinct, 7 unique	91 distinct, 9 unique	186 distinct, 4 unique
1	307	Store Services (92)	1	27
2	136	Lotto (271)	2	100
3	307	Store Services (92)	1	274
4	307	Store Services (92)	1	172
5	136	Lotto (271)	2	420
6	307	Store Services (92)	1	156
7	307	Store Services (92)	1	298
8	136	Lotto (271)	2	190
9				

Column statistics

Count	2991
Error	0
Empty	0
Distinct	60
Unique	7
Empty string	0
Min	Automot...
Max	Wine

Value distribution

We can create a new column to group some of these items



Groups

Name: super_parent_name

Field: parent_name

Group type: List

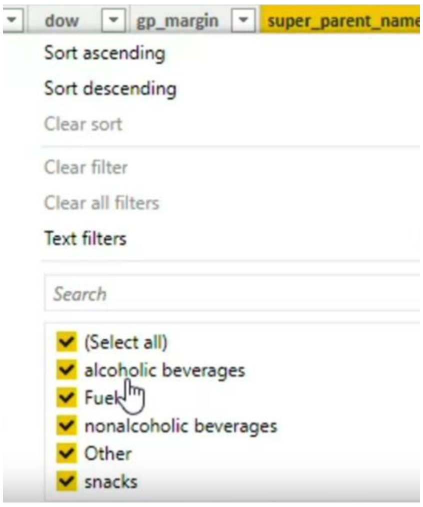
Ungrouped values

- Eggs/cheese
- Energy carbonated
- Energy non-carb
- Food-
- Food-Grocery
- Frozen Dispensed Beverage
- Frozen Foods
- Fuel
- Gift Cards - SKU (938)
- Gm/novelty

Groups and members

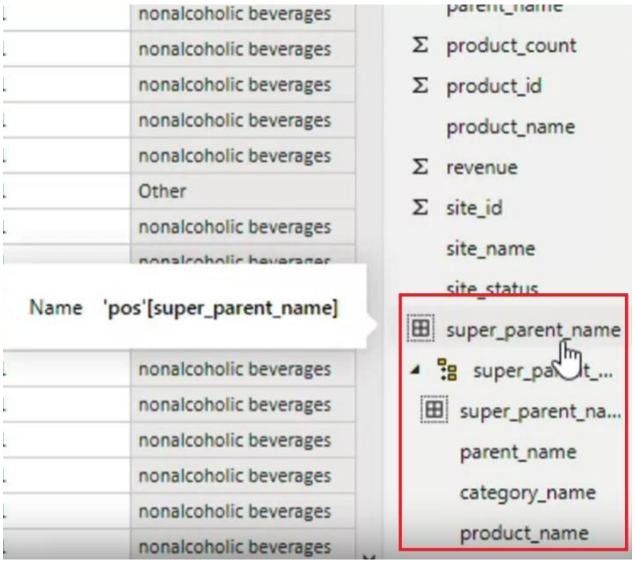
- alcoholic beverages
  - Beer
  - Liquor
  - Wine

Now we have just 5 different items



☑ (Select all)
☑ alcoholic beverages
☑ Fuel
☑ nonalcoholic beverages
☑ Other
☑ snacks

We can create a new hierarchy for our new column



nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
Other
nonalcoholic beverages
nonalcoholic beverages

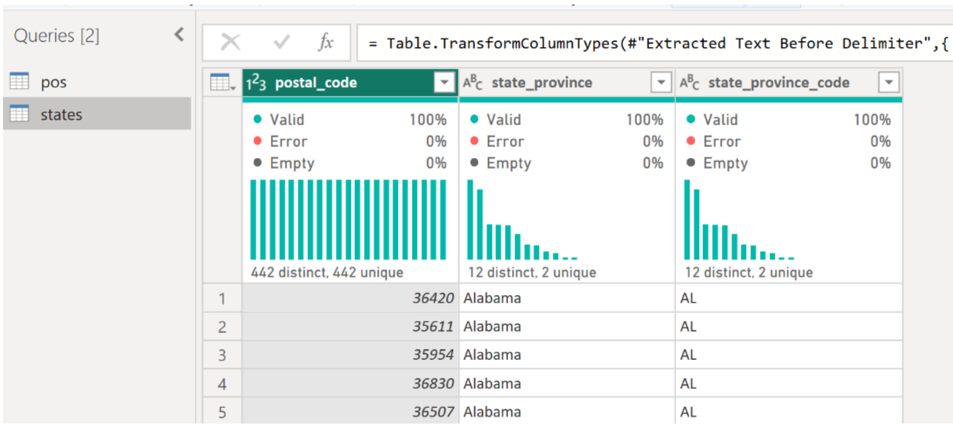
Name 'pos'[super_parent_name]

nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages
nonalcoholic beverages

ETL 4 : Data Models and Joins

Load up dataset states.csv that contains a list of states and join with the pos.csv dataset.

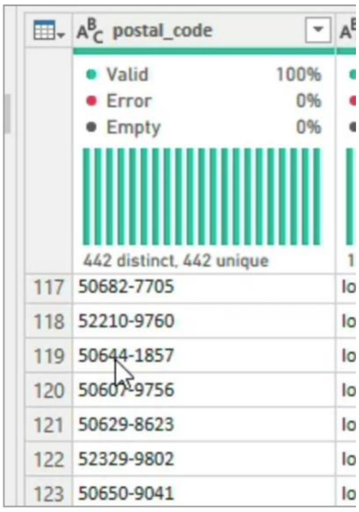
Note that the states dataset does not have a zip code column. It is instead called postal_code



```
= Table.TransformColumnTypes(#"Extracted Text Before Delimiter",{
```

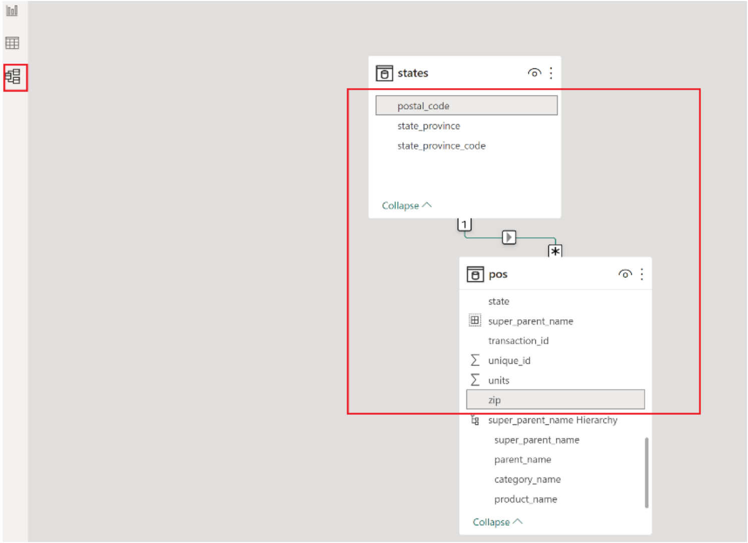
	postal_code	state_province	state_province_code
	Valid 100% Error 0% Empty 0% 442 distinct, 442 unique	Valid 100% Error 0% Empty 0% 12 distinct, 2 unique	Valid 100% Error 0% Empty 0% 12 distinct, 2 unique
1	36420	Alabama	AL
2	35611	Alabama	AL
3	35954	Alabama	AL
4	36830	Alabama	AL
5	36507	Alabama	AL

However, some of the values contain a 9-digit postal_code



	postal_code
	Valid 100% Error 0% Empty 0% 442 distinct, 442 unique
117	50682-7705
118	52210-9760
119	50644-1857
120	50607-9756
121	50629-8623
122	52329-9802
123	50650-9041

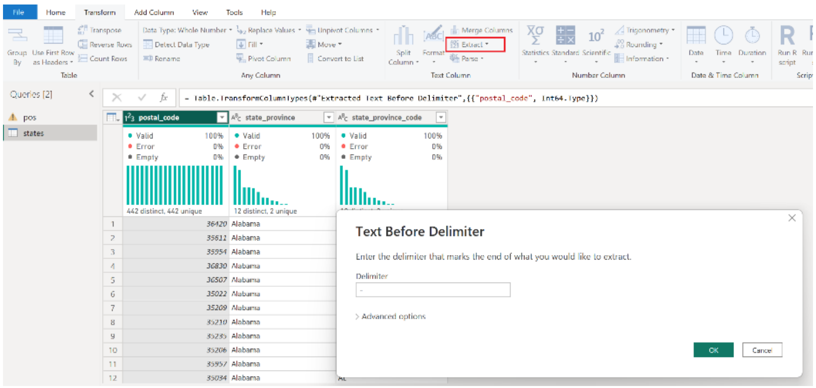
Go to the Model View and create a connection between the two datasets



We can see that there is a One-To-Many relationship between the states and pos table.

Each transaction in pos matches up with only one observation in the states table. But one entry in the states table could match up with many entries in the pos table.

To remove the last 4 digits of the post codes with 9 digits. Highlight the column, click on ‘Transform’, then ‘Extract’, ‘Text Before Delimiter’ and specify – (hyphen)



```
= Table.TransformColumnTypes(#"Extracted Text Before Delimiter",{{"postal_code", Int64.Type}})
```

	postal_code	state_province	state_province_code
	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
	442 distinct, 442 unique	12 distinct, 2 unique	
1	36420	Alabama	
2	35611	Alabama	
3	35954	Alabama	
4	36830	Alabama	
5	36507	Alabama	
6	35022	Alabama	
7	35209	Alabama	
8	35210	Alabama	
9	35235	Alabama	
10	35206	Alabama	
11	35957	Alabama	
12	35034	Alabama	AL

Text Before Delimiter

Enter the delimiter that marks the end of what you would like to extract.

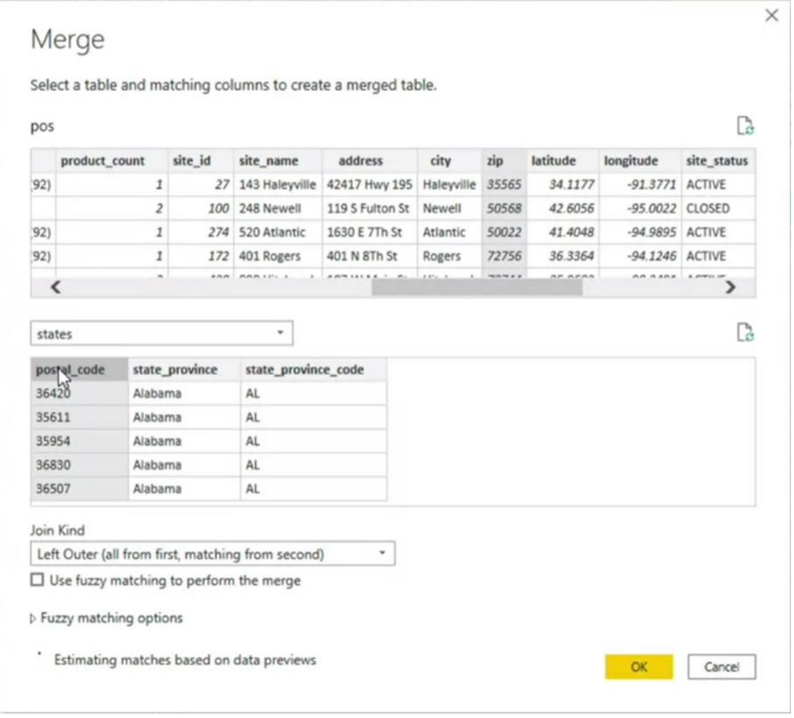
Delimiter: -

› Advanced options

OK Cancel

Use Merge Queries to link the zip from the pos dataset to the postal_code of the states dataset.

Note that the columns must be the same type.



Merge

Select a table and matching columns to create a merged table.

pos

	product_count	site_id	site_name	address	city	zip	latitude	longitude	site_status
92)	1	27	143 Haleyville	42417 Hwy 195	Haleyville	35565	34.1177	-91.3771	ACTIVE
	2	100	248 Newell	119 S Fulton St	Newell	50568	42.6056	-95.0022	CLOSED
92)	1	274	520 Atlantic	1630 E 7Th St	Atlantic	50022	41.4048	-94.9895	ACTIVE
92)	1	172	401 Rogers	401 N 8Th St	Rogers	72756	36.3364	-94.1246	ACTIVE

states

postal_code	state_province	state_province_code
36420	Alabama	AL
35611	Alabama	AL
35954	Alabama	AL
36830	Alabama	AL
36507	Alabama	AL

Join Kind: Left Outer (all from first, matching from second)

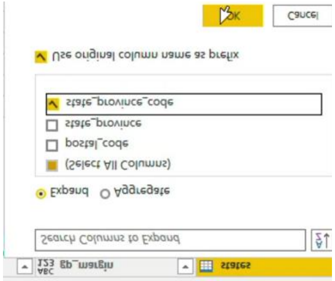
□ Use fuzzy matching to perform the merge

▷ Fuzzy matching options

Estimating matches based on data previews

OK Cancel

We only want to keep the state_province_code column



GP_margin ▲ states

Search Columns to Expand

● Expand ○ Aggregate

▨ (Select All Columns)

□ postal_code

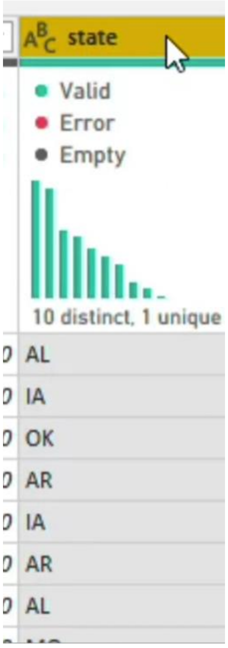
□ state_province

☑ state_province_code

☑ Use original column name as prefix

OK Cancel

The column name is a bit long, so rename it to state



```
10 distinct, 1 unique
AL
IA
OK
AR
IA
AR
AL
```

Use star schema, one main dataset, and any other dataset is joined to that key dataset to avoid querying lots of hierarchical relationships.