

Project1_Customers_Messy

Source: Where might the data have come from?

Corporate office, their system and database of the customers that are registered online.

Format: Are the columns structured consistently?

The columns are structured consistently

Quality: Identify missing values, duplicates, invalid entries.

CustomerID 1006 - has a blank first name

CustomerID 1011 - has a -1 as an age

CustomerID 1016 - has an Unknown Gender

CustomerID 1021 - has misspelled city Sea_ttle

1	CustomerID	FirstName	LastName	Gender	Age	City	State	Segment
2	1001	Michael	Lopez	F	33	Tacoma	WA	Retail
3	1002	Michael	Thompson	M	55	Bellevue	WA	Retail
4	1003	Sarah	Chen	M	32	Olympia	WA	Online
5	1004	Sarah	Thompson	M	63	Olympia	WA	Wholesale
6	1005	David	White	F	69	Seattle	WA	Retail
7	1006		Taylor	F	27	Tacoma	WA	Wholesale
8	1007	Michael	Chen	F	24	Eugene	OR	Wholesale
9	1008	Harper	Wilson	M	64	Vancouver	WA	Online
10	1009	Michael	Thomas	M	53	Salem	OR	Online
11	1010	Harper	Anderson	M	63	Portland	OR	Retail
12	1011	David	Moore	M	-1	Portland	OR	Wholesale
13	1012	Emma	White	F	28	Eugene	OR	Wholesale
14	1013	Sophia	Wilson	M	56	Spokane	WA	Online
15	1014	David	Brown	F	42	Salem	OR	Online
16	1015	Amelia	Miller	F	67	Seattle	WA	Retail
17	1016	James	Taylor	Unknown	35	Portland	OR	Retail
18	1017	Lucas	Taylor	M	59	Vancouver	WA	Wholesale
19	1018	Ethan	Johnson	F	26	Tacoma	WA	Online
20	1019	Amelia	Thompson	F	65	Bend	OR	Wholesale
21	1020	Lucas	Thomas	F	32	Spokane	WA	Online
22	1021	Charlotte	Chen	M	25	Sea_ttle	WA	Online
23	1022	Daniel	Jackson	M	42	Bellevue	WA	Online
24	1023	Ethan	Martin	F	53	Seattle	WA	Online
25	1024	Michael	Thompson	F	67	Eugene	OR	Retail
26	1025	John	Jackson	M	47	Seattle	WA	Online

Risks: What problems would these issues cause in analysis?

Having a missing value can cause errors in aggregation and distort counts in queries causing inaccurate analysis. Mistype data values like in CustID 1021 can cause incorrect count during query.

Project1_Customers_Messy

Source: Where might the data have come from?

It's a transactional sales database that could come from a point-of-sale (POS) system in retail or e-commerce. Or an ERP/CRM where customer, product, and sales record.

Format: Are the columns structured consistently?

The table's column structure looks good and the data stays in its column.

Quality: Identify missing values, duplicates, invalid entries.

SaleID 3006 - Date is shifted to the left causing the format of the whole table to be misaligned.

SaleID 3011 - Negative quantity which doesn't make sense.

SaleID 3016 - CustomerID is 99999 which is irregular from the rest of the CustomerID.

SaleID 3021 - ProductID is 99999 similar issue to the previous error.

SaleID 3026 - TotalAmount is spelled out rather than a numerical value.

Risks: What problems would these issues cause in analysis?

The bad input makes it harder for accurate data when the data is being queried.

Having invalid dates like “2023-13-40” can cause errors when sorting. Negative values in Quantity such as “-3” can distort sales totals. Using placeholder values like “99999” in CustomerID or ProductID can cause misleading results in analysis.

Project1_Sales_Messy

Source: Where might the data have come from?

The data likely comes from a company’s product catalog or inventory database that tracks details, costs, and suppliers.

Format: Are the columns structured consistently?

The columns are mostly structured consistently with ProductID, ProductName, Category, UnitPrice, Cost, and Supplier, but have bad data.

Quality: Identify missing values, duplicates, invalid entries.

ProductID - 2003 has a missing UnitPrice

ProductID - 2005 has a missing Supplier

ProductID - 2002 Is a Duplicate to Row 2

Risks: What problems would these issues cause in analysis?

Missing or invalid values can cause errors in cost or profit calculations. Duplicates causes misinformation leading to distort financial analysis.

Attempt to Import

Customer Messy

Import Flat File 'Project1_ClientData'

Preview Data

Introduction Help

Specify Input File

Preview Data

Modify Columns

Summary

Results

Preview Data

This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.

ID	FName	LName	Sex	Years	Town	Reg
1001	Michael	Lopez	F	33	Tacoma	WA
1002	Michael	Thompson	M	55	Bellevue	WA
1003	Sarah	Chen	M	32	Olympia	WA
1004	Sarah	Thompson	M	63	Olympia	WA
1005	David	White	F	69	Seattle	WA
1006		Taylor	F	27	Tacoma	WA
1007	Michael	Chen	F	24	Eugene	OR
1008	Harper	Wilson	M	64	Vancouver	WA
1009	Michael	Thomas	M	53	Salem	OR
1010	Harper	Anderson	M	63	Portland	OR
1011	David	Moore	M	-1	Portland	OR
1012	Emma	White	F	28	Eugene	OR
1013	Sophia	Wilson	M	56	Spokane	WA
1014	David	Brown	F	42	Salem	OR
1015	Amelia	Miller	F	67	Seattle	WA
1016	James	Taylor	Unknown	35	Portland	OR
1017	Lucas	Taylor	M	59	Vancouver	WA
1018	Ethan	Johnson	F	26	Tacoma	WA
1019	Amelia	Thompson	F	65	Bend	OR
1020	Lucas	Thomas	F	32	Spokane	WA

Use Rich Data Type Detection - may provide a closer type fit. However, cells with anomalous values may be dropped.

< Previous Next > Cancel

System doesn't catch blanks or unknown

Import Flat File 'Project1_ClientData'

Modify Columns

Introduction Help

Specify Input File

Preview Data

Modify Columns

Summary

Results

Modify Columns

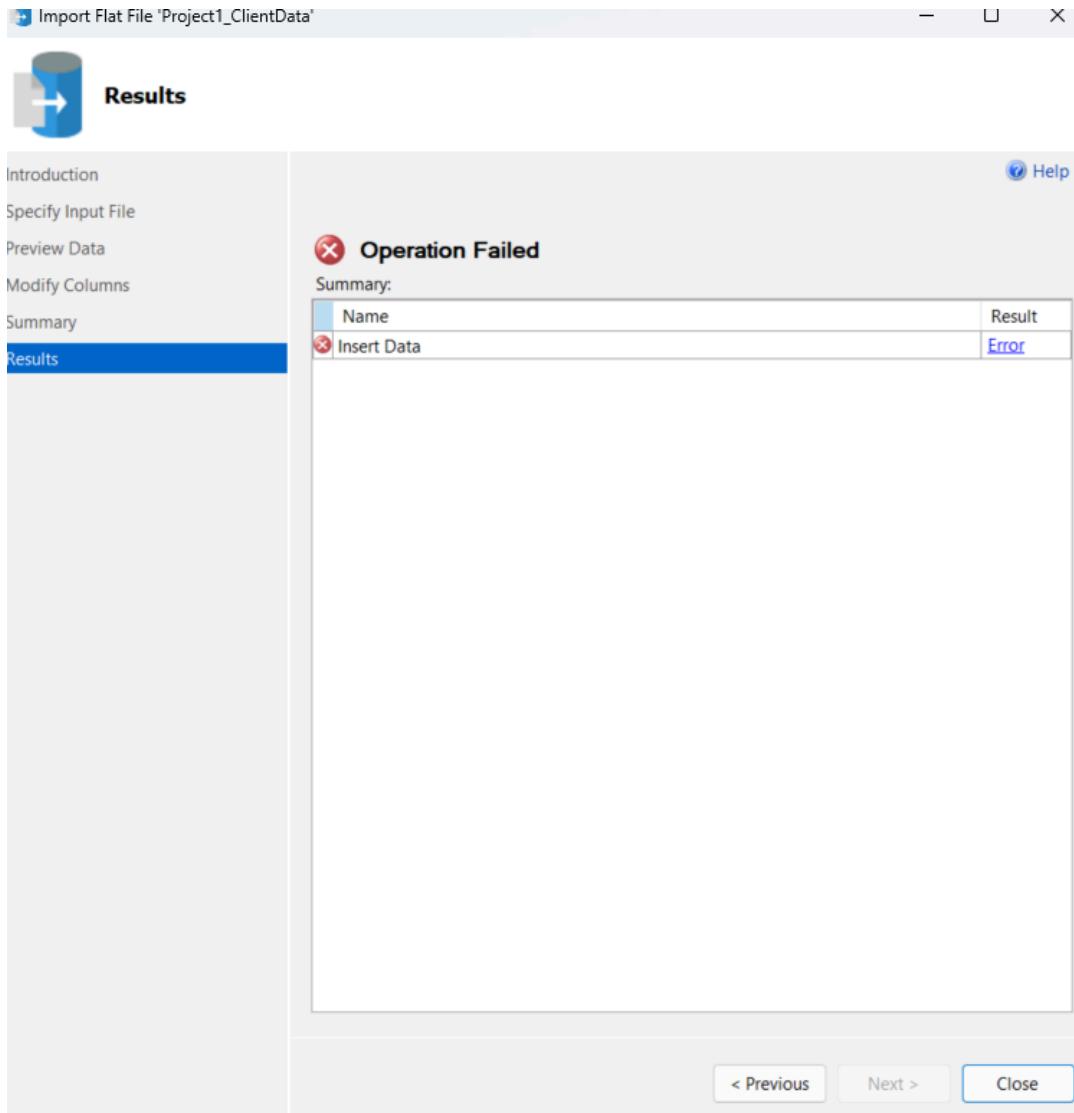
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	<input type="checkbox"/> Allow Nulls
ID	smallint	<input checked="" type="checkbox"/>	<input type="checkbox"/>
FName	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
LName	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Sex	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Years	smallint	<input type="checkbox"/>	<input type="checkbox"/>
Town	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Region	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Group	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>

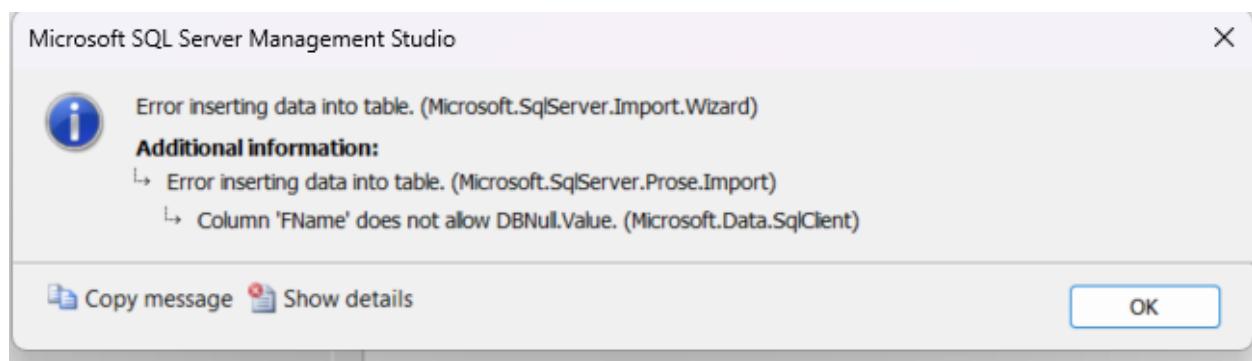
Row granularity of error reporting (performance impact with smaller ranges)

< Previous Next > Cancel

System saw the blank space and auto nullable the missing first name



After clicking through its throwing an error



Product Messy

Import Flat File 'Project1_ClientData'

Preview Data

Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results

Help

Preview Data

This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.

ProductID	ProductName	Category	UnitPrice	Cost	Supplier
2001	Coffee Bean...	Beverages	12.99	7.5	Pacific Roas...
2002	Tea Sample...	Beverages	18.99	10.0	Leaf & Co
2003	Espresso M...	Equipment	N/A	180.0	BrewTech
2004	French Press	Equipment	39.99	20.0	CafeGear
2005	Mug Set	Merchandise	24.99	10.0	
2006	Cold Brew Kit	Equipment	59.99	35.0	CoolBrew
2007	Green Tea 1...	Beverages	9.99	4.0	Leaf & Co
2008	Herbal Tea ...	Beverages	14.99	7.0	Leaf & Co
2009	Coffee Grin...	Equipment	79.99	50.0	BrewTech
2010	Tumbler	Merchandise	19.99	8.0	KitchenWorks
2002	Tea Sample...	Beverages	18.99	10.0	Leaf & Co

Use Rich Data Type Detection - may provide a closer type fit. However, cells with anomalous values may be dropped.

< Previous Next > Cancel

It doesn't auto fill with something, just let it be null

Import Flat File 'Project1_ClientData'

Modify Columns

Introduction Specify Input File Preview Data Modify Columns Summary Results

Help

Modify Columns

This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	Allow Nulls
ProductID	smallint	<input type="checkbox"/>	<input type="checkbox"/>
ProductName	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Category	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
UnitPrice	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Cost	float	<input type="checkbox"/>	<input type="checkbox"/>
Supplier	nvarchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Row granularity of error reporting (performance impact with smaller ranges) No Range

< Previous Next > Cancel

Then enforce rules to make it work

Sales Messy

Import Flat File 'Project1_ClientData'

 Preview Data

Introduction
Specify Input File
Preview Data
Modify Columns
Summary
Results

Help

Preview Data

This operation analyzed the input file structure to generate the preview below for up to the first 50 rows.

SaleID	CustomerID	ProductID	Date	Quantity	TotalAmount
3016	99999	2007	2023-05-27	1	9.99
3017	1061	2001	2023-12-18	1	12.99
3018	1045	2004	2023-11-03	1	39.99
3019	1097	2001	2023-04-07	1	12.99
3020	1080	2003	2023-04-05	4	999.96
3021	1086	99999	2023-10-07	4	75.96
3022	1090	2005	2023-06-06	5	124.95
3023	1078	2002	2023-03-10	1	18.99
3024	1075	2001	2023-05-19	4	51.96
3025	1051	2004	2023-02-19	2	79.98
3026	1014	2005	2023-11-20	1	fifty
3027	1073	2001	2023-06-18	4	51.96
3028	1085	2006	2023-02-17	3	179.97
3029	1002	2007	2023-08-04	4	39.96
3030	1047	2008	2023-12-05	4	59.96
3031	1023	2009	2023-11-09	5	399.95
3032	1069	2008	2023-08-14	5	74.95
3033	1035	2006	2023-04-27	1	59.99
3034	1036	2008	2023-04-25	4	59.96
3035	1073	2010	2023-11-13	3	59.97
3036	1004	2008	2023-06-06	4	59.96

Use Rich Data Type Detection - may provide a closer type fit. However, cells with anomalous values may be dropped.

< Previous Next > Cancel

Import Flat File 'Project1_ClientData'

Modify Columns

Introduction Specify Input File Preview Data Modify Columns Summary Results

Help

Modify Columns

This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	Allow Nulls
SaleID	smallint	<input checked="" type="checkbox"/>	<input type="checkbox"/>
CustomerID	int	<input type="checkbox"/>	<input type="checkbox"/>
ProductID	int	<input type="checkbox"/>	<input type="checkbox"/>
Date	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Quantity	smallint	<input type="checkbox"/>	<input type="checkbox"/>
TotalAmount	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Row granularity of error reporting (performance impact with smaller ranges) No Range

< Previous Next > Cancel

It caught the miss data type in the column, and then made the column nullable to compensate.

All attempts to import without correction ends in an error.

Reflection: Client data often contains inconsistencies like missing values, duplicates, and invalid formats all from human error that is very hard to avoid completely. Even small errors can create big problems in analysis, so cleaning and validating data is just as important as the analysis itself.