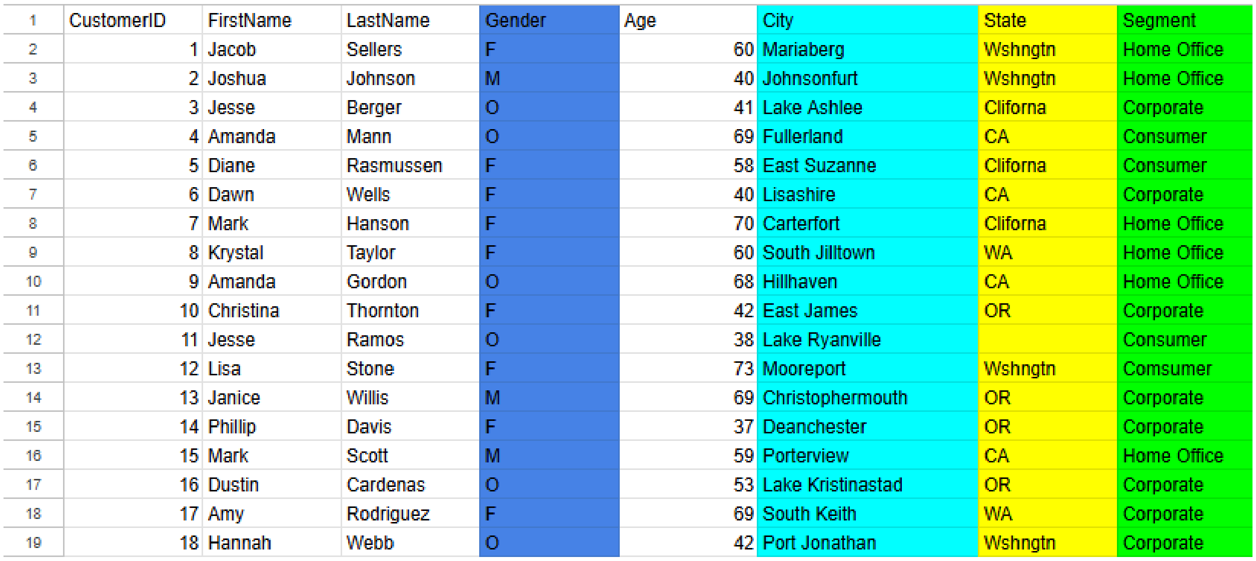
Capstone Project Wk2: ETL Package Deliverable

Customers Data



1	CustomerID	FirstName	LastName	Gender	Age	City	State	Segment
2	1	Jacob	Sellers	F	60	Mariaberg	Wshngtn	Home Office
3	2	Joshua	Johnson	M	40	Johnsonfurt	Wshngtn	Home Office
4	3	Jesse	Berger	O	41	Lake Ashlee	Cliforna	Corporate
5	4	Amanda	Mann	O	69	Fullerland	CA	Consumer
6	5	Diane	Rasmussen	F	58	East Suzanne	Cliforna	Consumer
7	6	Dawn	Wells	F	40	Lisashire	CA	Corporate
8	7	Mark	Hanson	F	70	Carterfort	Cliforna	Home Office
9	8	Krystal	Taylor	F	60	South Jilltown	WA	Home Office
10	9	Amanda	Gordon	O	68	Hillhaven	CA	Home Office
11	10	Christina	Thornton	F	42	East James	OR	Corporate
12	11	Jesse	Ramos	O	38	Lake Ryanville		Consumer
13	12	Lisa	Stone	F	73	Mooreport	Wshngtn	Comsumer
14	13	Janice	Willis	M	69	Christophermouth	OR	Corporate
15	14	Phillip	Davis	F	37	Deanchester	OR	Corporate
16	15	Mark	Scott	M	59	Porterview	CA	Home Office
17	16	Dustin	Cardenas	O	53	Lake Kristinastad	OR	Corporate
18	17	Amy	Rodriguez	F	69	South Keith	WA	Corporate
19	18	Hannah	Webb	O	42	Port Jonathan	Wshngtn	Corporate

Issues

1. Some genders have an O
2. Cities names structures seems to be missing spaces
    a. Capture every possible end of these cities that make sense, and then put a space in front of it and cap the first letter.
        1. berg, borough, burgh, fort, furt, port, side, land, ville, view, mouth, haven, shire, stad, town
3. State's
    a. Inconsistent format

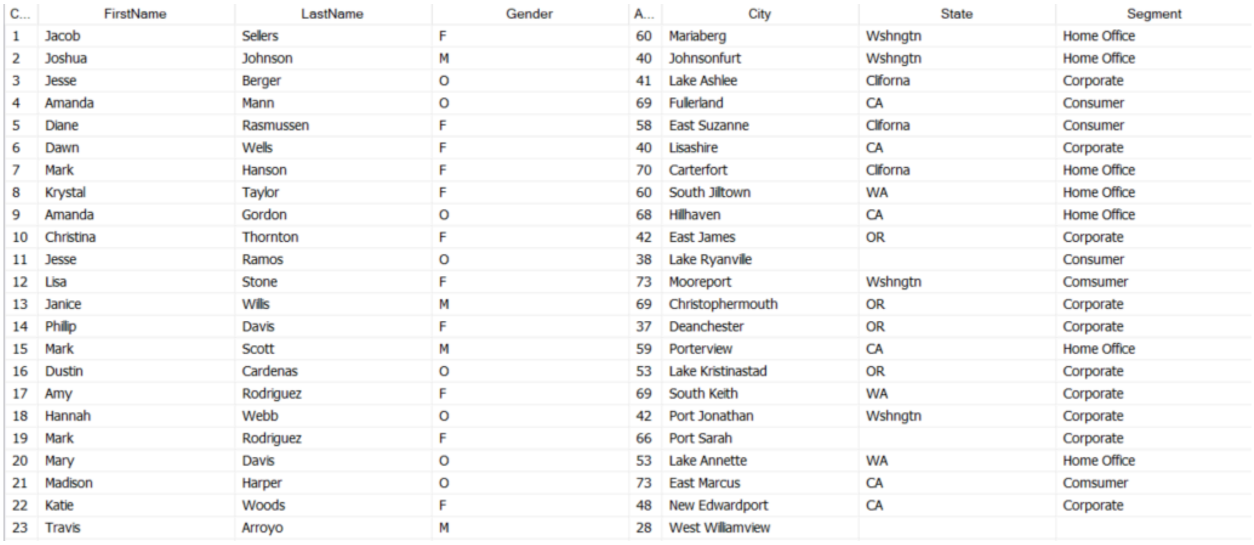
b. Spelling errors

4. Category

   a. Misspelled

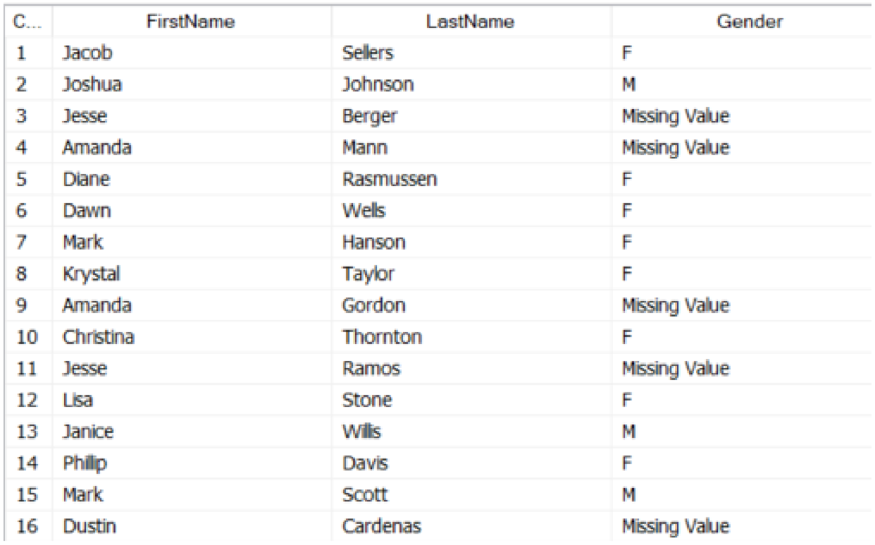
   b. Null spaces

Process of Fixing the Issues



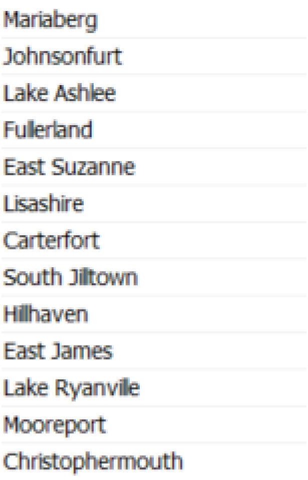
C...	FirstName	LastName	Gender	A...	City	State	Segment
1	Jacob	Sellers	F	60	Mariaberg	Wshngtn	Home Office
2	Joshua	Johnson	M	40	Johnsonfurt	Wshngtn	Home Office
3	Jesse	Berger	O	41	Lake Ashlee	Clforna	Corporate
4	Amanda	Mann	O	69	Fullerland	CA	Consumer
5	Diane	Rasmussen	F	58	East Suzanne	Clforna	Consumer
6	Dawn	Wells	F	40	Lisashire	CA	Corporate
7	Mark	Hanson	F	70	Carterfort	Clforna	Home Office
8	Krystal	Taylor	F	60	South Jiltown	WA	Home Office
9	Amanda	Gordon	O	68	Hillhaven	CA	Home Office
10	Christina	Thornton	F	42	East James	OR	Corporate
11	Jesse	Ramos	O	38	Lake Ryanville		Consumer
12	Lisa	Stone	F	73	Mooreport	Wshngtn	Comsumer
13	Janice	Wills	M	69	Christophermouth	OR	Corporate
14	Phillip	Davis	F	37	Deanchester	OR	Corporate
15	Mark	Scott	M	59	Porterview	CA	Home Office
16	Dustin	Cardenas	O	53	Lake Kristinastad	OR	Corporate
17	Amy	Rodriguez	F	69	South Keith	WA	Corporate
18	Hannah	Webb	O	42	Port Jonathan	Wshngtn	Corporate
19	Mark	Rodriguez	F	66	Port Sarah		Corporate
20	Mary	Davis	O	53	Lake Annette	WA	Home Office
21	Madison	Harper	O	73	East Marcus	CA	Comsumer
22	Katie	Woods	F	48	New Edwardport	CA	Corporate
23	Travis	Arroyo	M	28	West Williamview		

Imported data from csv files



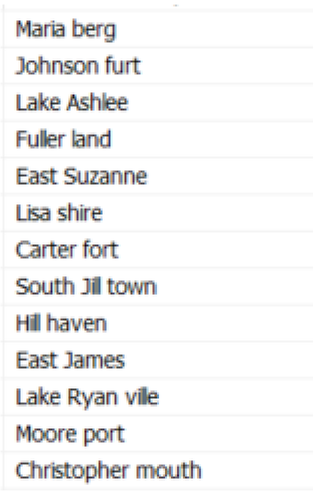
C...	FirstName	LastName	Gender
1	Jacob	Sellers	F
2	Joshua	Johnson	M
3	Jesse	Berger	Missing Value
4	Amanda	Mann	Missing Value
5	Diane	Rasmussen	F
6	Dawn	Wells	F
7	Mark	Hanson	F
8	Krystal	Taylor	F
9	Amanda	Gordon	Missing Value
10	Christina	Thornton	F
11	Jesse	Ramos	Missing Value
12	Lisa	Stone	F
13	Janice	Willis	M
14	Phillip	Davis	F
15	Mark	Scott	M
16	Dustin	Cardenas	Missing Value

Changed 0 in gender to missing value



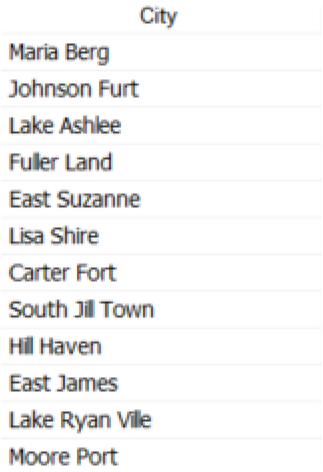
Mariaberg
Johnsonfurt
Lake Ashlee
Fullerland
East Suzanne
Lisashire
Carterfort
South Jiltown
Hilhaven
East James
Lake Ryanvile
Mooreport
Christophermouth

==>



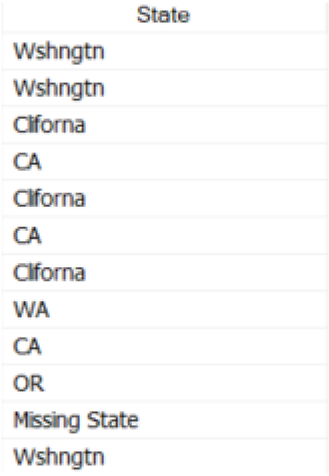
Maria berg
Johnson furt
Lake Ashlee
Fuller land
East Suzanne
Lisa shire
Carter fort
South Jil town
Hil haven
East James
Lake Ryan vile
Moore port
Christopher mouth

==>



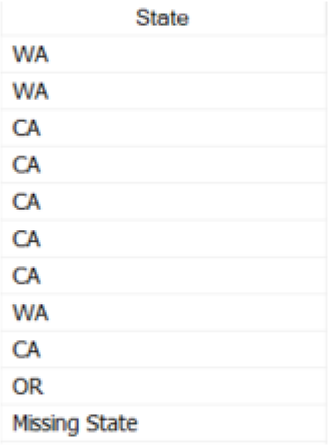
City
Maria Berg
Johnson Furt
Lake Ashlee
Fuller Land
East Suzanne
Lisa Shire
Carter Fort
South Jil Town
Hil Haven
East James
Lake Ryan Vile
Moore Port

Splits suffices from cities names then capitalize the first letter after a space



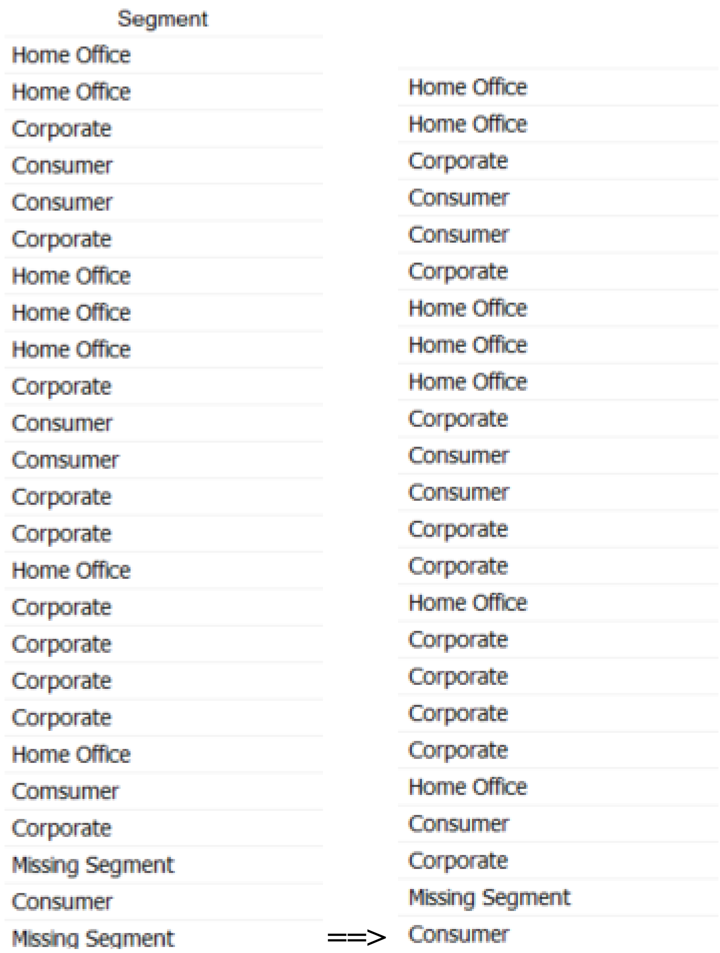
State
Wshngtn
Wshngtn
Clforna
CA
Clforna
CA
Clforna
WA
CA
OR
Missing State
Wshngtn

==>



State
WA
WA
CA
CA
CA
CA
CA
WA
CA
OR
Missing State

Fills null with "Missing States" then insert proper abbreviated states.



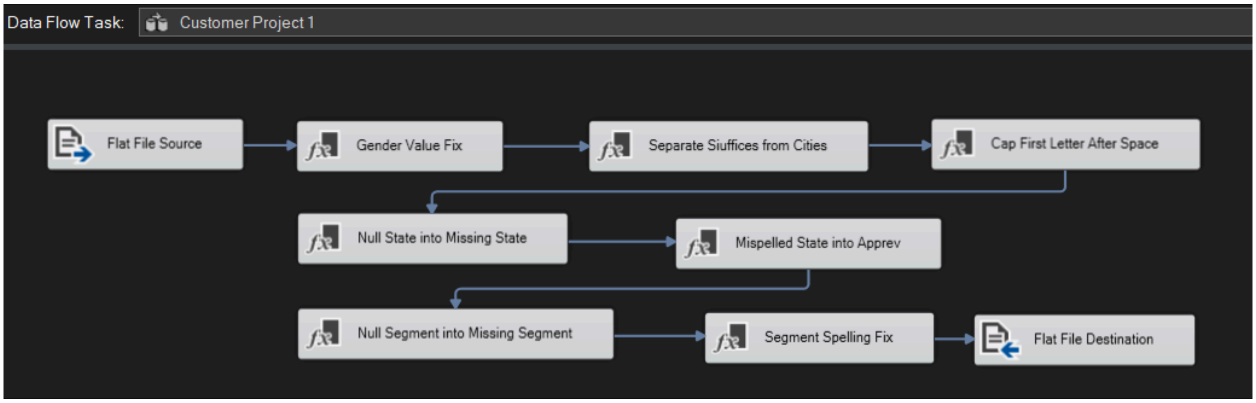
Segment
Home Office
Home Office
Corporate
Consumer
Consumer
Corporate
Home Office
Home Office
Home Office
Corporate
Consumer
Comsumer
Corporate
Corporate
Home Office
Corporate
Corporate
Corporate
Corporate
Home Office
Comsumer
Corporate
Missing Segment
Consumer
Missing Segment

==>

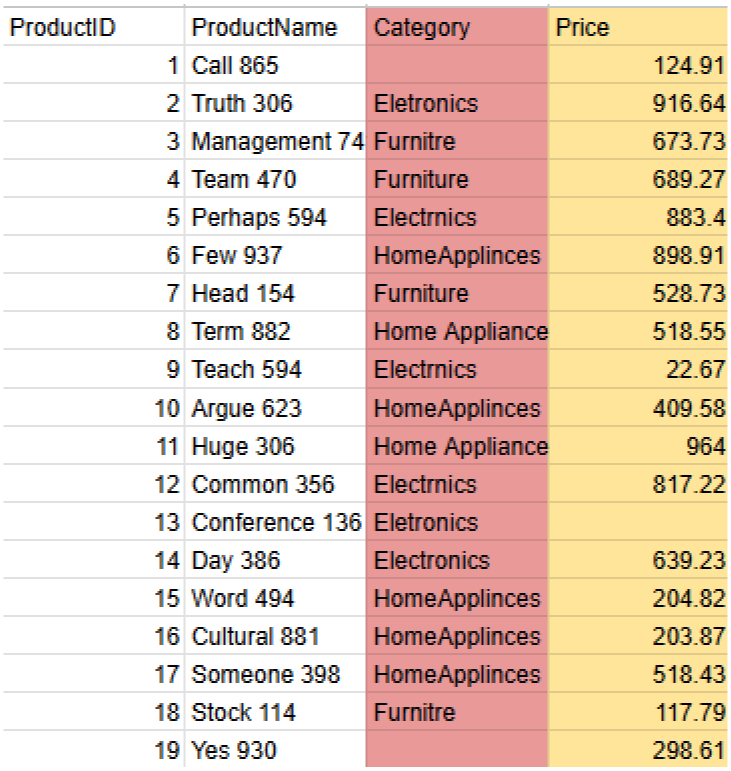
Home Office
Home Office
Corporate
Consumer
Consumer
Corporate
Home Office
Home Office
Home Office
Corporate
Consumer
Consumer
Corporate
Corporate
Home Office
Corporate
Corporate
Corporate
Corporate
Home Office
Consumer
Corporate
Missing Segment
Consumer

Fill in with "Missing Segment" and then correct misspelled Segments

Customer Data Flow


Data Flow Task:
Customer Project 1
Flat File Source
Gender Value Fix
Separate Siuffices from Cities
Cap First Letter After Space
Null State into Missing State
Mispelled State into Apprev
Null Segment into Missing Segment
Segment Spelling Fix
Flat File Destination

Products Data

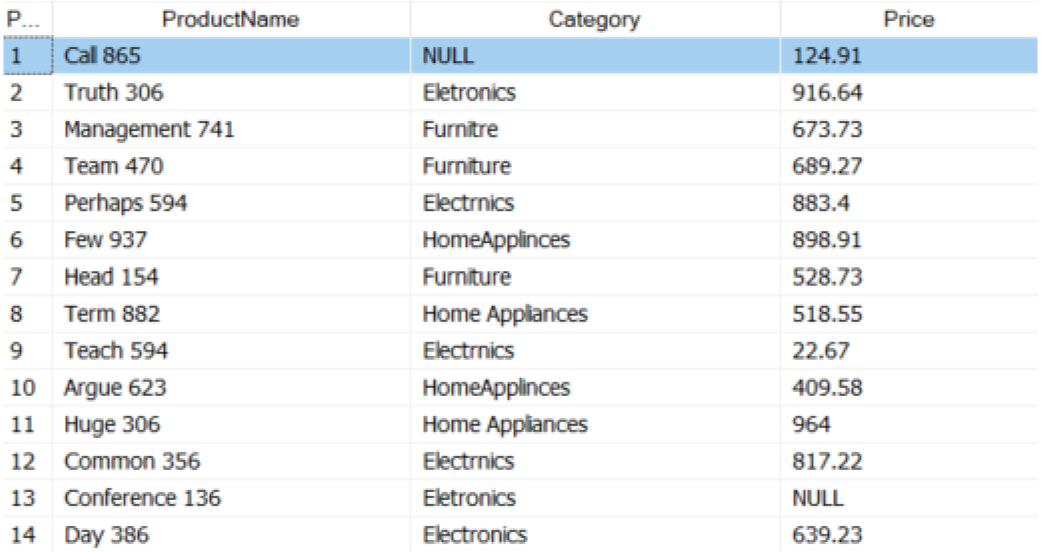


ProductID	ProductName	Category	Price
1	Call 865		124.91
2	Truth 306	Eletronics	916.64
3	Management 74	Furnitre	673.73
4	Team 470	Furniture	689.27
5	Perhaps 594	Electrnics	883.4
6	Few 937	HomeApplinces	898.91
7	Head 154	Furniture	528.73
8	Term 882	Home Appliance	518.55
9	Teach 594	Electrnics	22.67
10	Argue 623	HomeApplinces	409.58
11	Huge 306	Home Appliance	964
12	Common 356	Electrnics	817.22
13	Conference 136	Eletronics	
14	Day 386	Electronics	639.23
15	Word 494	HomeApplinces	204.82
16	Cultural 881	HomeApplinces	203.87
17	Someone 398	HomeApplinces	518.43
18	Stock 114	Furnitre	117.79
19	Yes 930		298.61

Issues

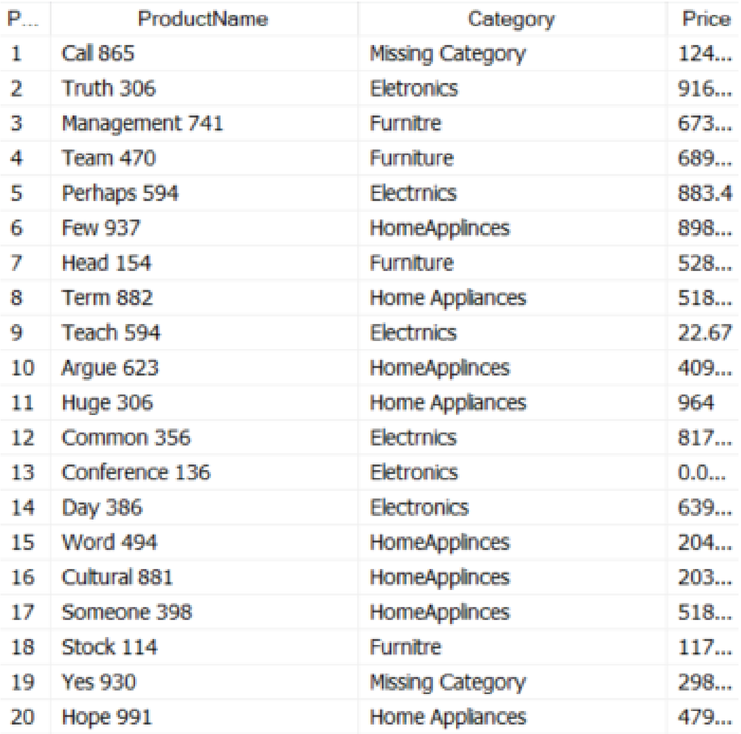
1. Category
   a. Null/Empty
   b. Mispell words
2. Price
   a. Missing Price

Process



P...	ProductName	Category	Price
1	Call 865	NULL	124.91
2	Truth 306	Eletronics	916.64
3	Management 741	Furnitre	673.73
4	Team 470	Furniture	689.27
5	Perhaps 594	Electrnics	883.4
6	Few 937	HomeApplinces	898.91
7	Head 154	Furniture	528.73
8	Term 882	Home Appliances	518.55
9	Teach 594	Electrnics	22.67
10	Argue 623	HomeApplinces	409.58
11	Huge 306	Home Appliances	964
12	Common 356	Electrnics	817.22
13	Conference 136	Eletronics	NULL
14	Day 386	Electronics	639.23

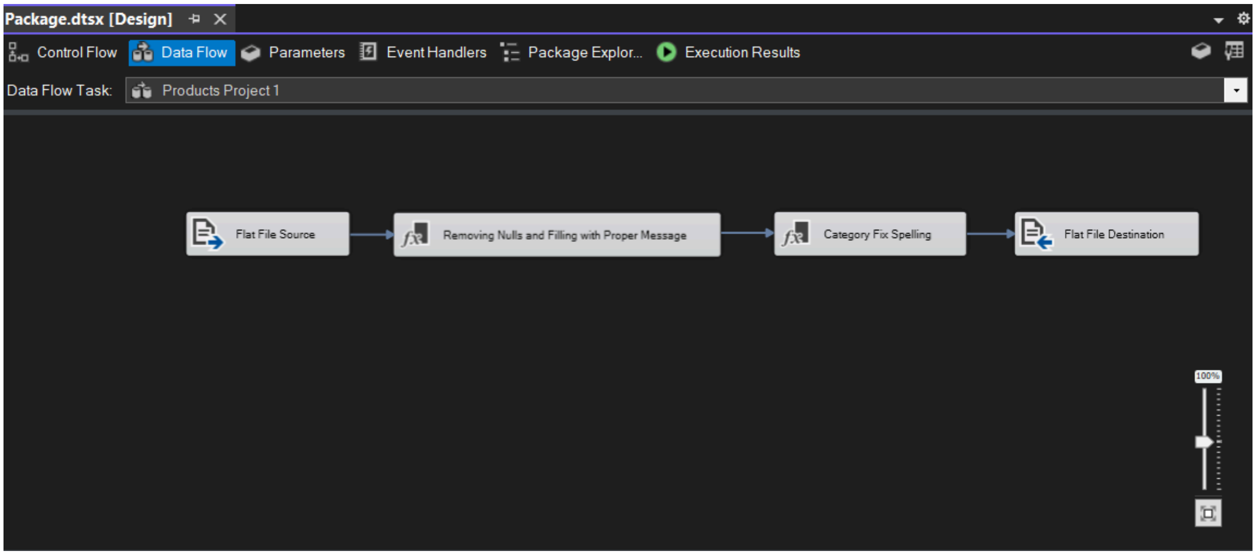
Removed Fill in Null/Empty with on Derived Column, since Derived Column can only mess with one column at a time.



P...	ProductName	Category	Price
1	Call 865	Missing Category	124...
2	Truth 306	Eletronics	916...
3	Management 741	Furnitre	673...
4	Team 470	Furniture	689...
5	Perhaps 594	Electrnics	883.4
6	Few 937	HomeApplinces	898...
7	Head 154	Furniture	528...
8	Term 882	Home Appliances	518...
9	Teach 594	Electrnics	22.67
10	Argue 623	HomeApplinces	409...
11	Huge 306	Home Appliances	964
12	Common 356	Electrnics	817...
13	Conference 136	Eletronics	0.0...
14	Day 386	Electronics	639...
15	Word 494	HomeApplinces	204...
16	Cultural 881	HomeApplinces	203...
17	Someone 398	HomeApplinces	518...
18	Stock 114	Furnitre	117...
19	Yes 930	Missing Category	298...
20	Hope 991	Home Appliances	479...

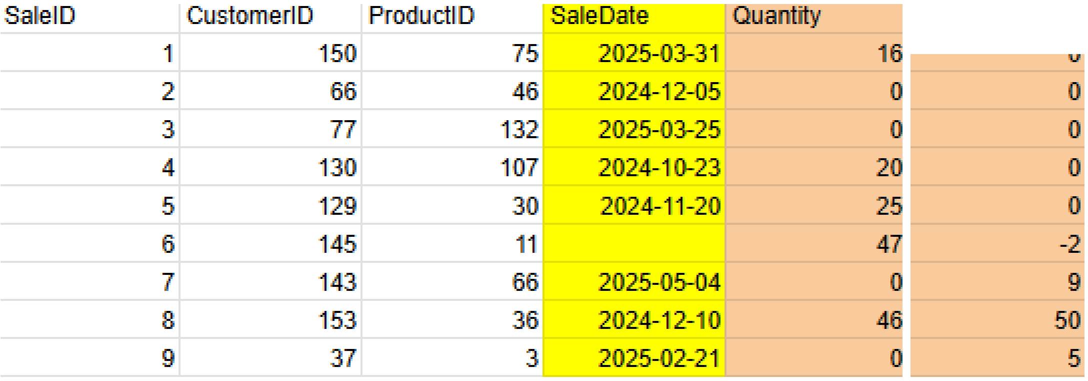
Fix Spelling

Data Flow for Products



Data Flow Task: Products Project 1

Sales Project



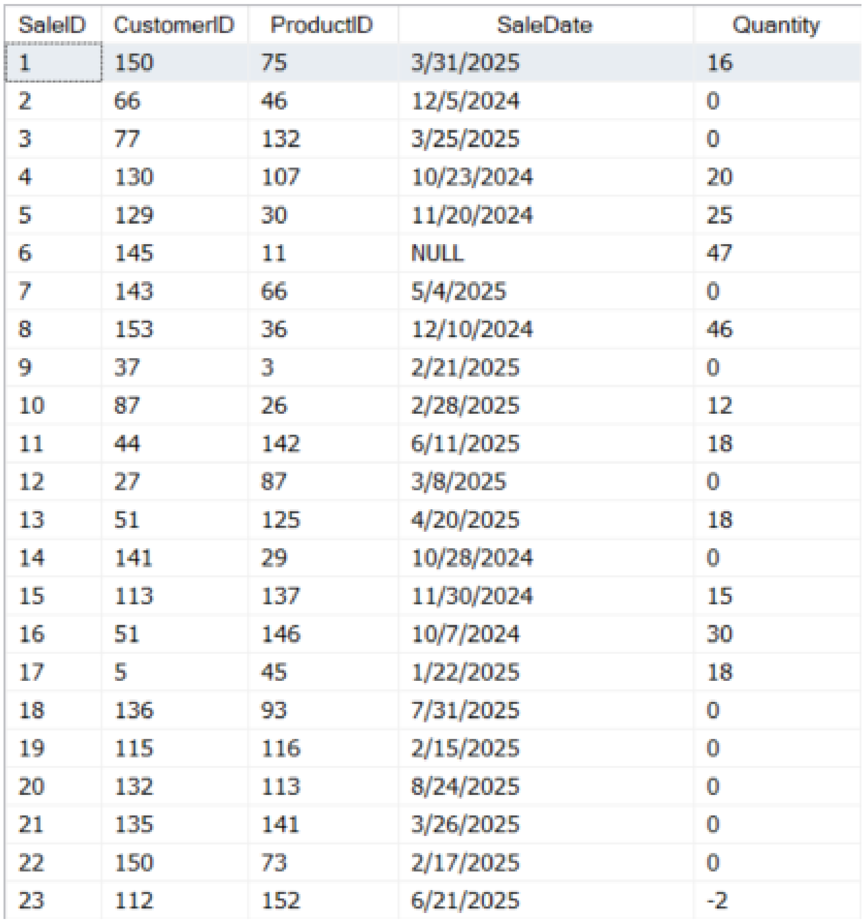
SaleID	CustomerID	ProductID	SaleDate	Quantity	
1	150	75	2025-03-31	16	0
2	66	46	2024-12-05	0	0
3	77	132	2025-03-25	0	0
4	130	107	2024-10-23	20	0
5	129	30	2024-11-20	25	0
6	145	11		47	-2
7	143	66	2025-05-04	0	9
8	153	36	2024-12-10	46	50
9	37	3	2025-02-21	0	5

Issues

1. SaleDate - Missing Dates

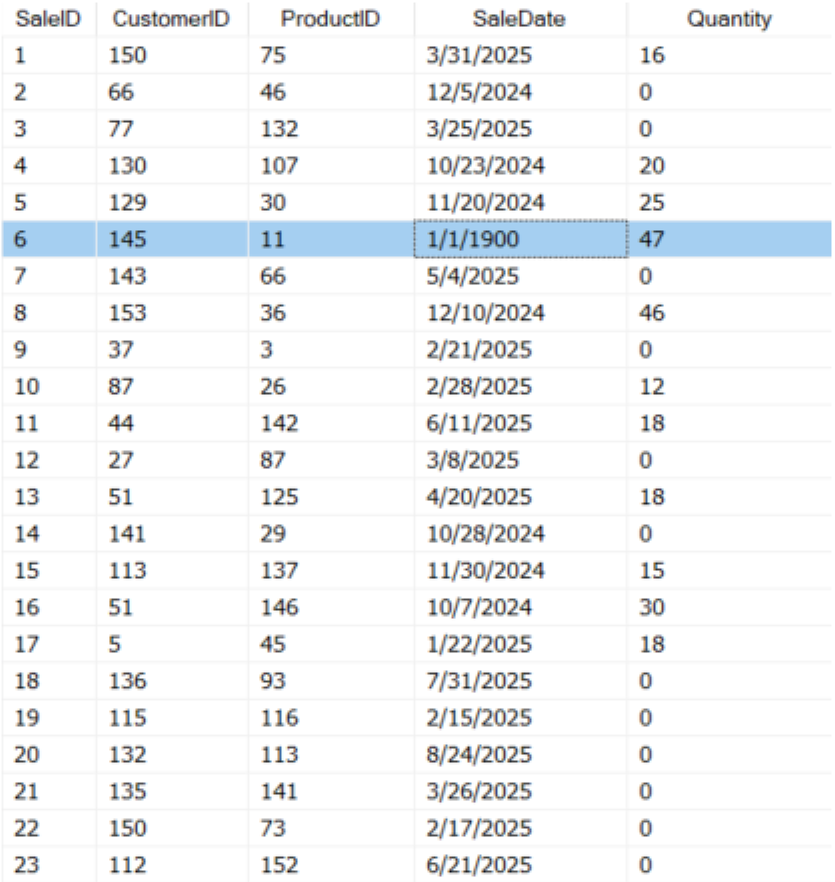
2. Quantity - Negative numbers

Process



SaleID	CustomerID	ProductID	SaleDate	Quantity
1	150	75	3/31/2025	16
2	66	46	12/5/2024	0
3	77	132	3/25/2025	0
4	130	107	10/23/2024	20
5	129	30	11/20/2024	25
6	145	11	NULL	47
7	143	66	5/4/2025	0
8	153	36	12/10/2024	46
9	37	3	2/21/2025	0
10	87	26	2/28/2025	12
11	44	142	6/11/2025	18
12	27	87	3/8/2025	0
13	51	125	4/20/2025	18
14	141	29	10/28/2024	0
15	113	137	11/30/2024	15
16	51	146	10/7/2024	30
17	5	45	1/22/2025	18
18	136	93	7/31/2025	0
19	115	116	2/15/2025	0
20	132	113	8/24/2025	0
21	135	141	3/26/2025	0
22	150	73	2/17/2025	0
23	112	152	6/21/2025	-2

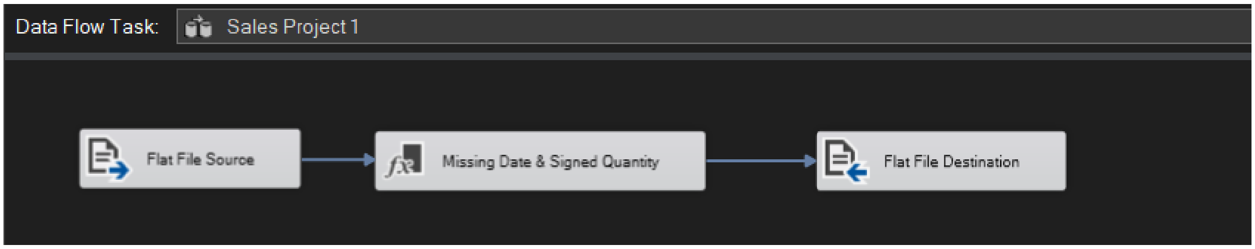
Import data to work with



SaleID	CustomerID	ProductID	SaleDate	Quantity
1	150	75	3/31/2025	16
2	66	46	12/5/2024	0
3	77	132	3/25/2025	0
4	130	107	10/23/2024	20
5	129	30	11/20/2024	25
6	145	11	1/1/1900	47
7	143	66	5/4/2025	0
8	153	36	12/10/2024	46
9	37	3	2/21/2025	0
10	87	26	2/28/2025	12
11	44	142	6/11/2025	18
12	27	87	3/8/2025	0
13	51	125	4/20/2025	18
14	141	29	10/28/2024	0
15	113	137	11/30/2024	15
16	51	146	10/7/2024	30
17	5	45	1/22/2025	18
18	136	93	7/31/2025	0
19	115	116	2/15/2025	0
20	132	113	8/24/2025	0
21	135	141	3/26/2025	0
22	150	73	2/17/2025	0
23	112	152	6/21/2025	0

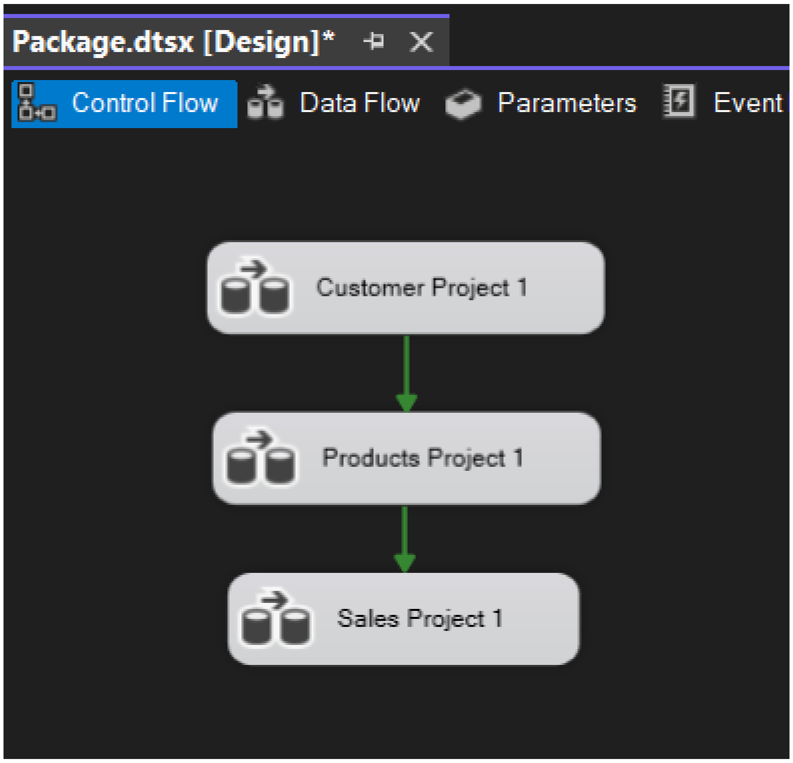
Replace empty date with a temp date, since it's going to just say Null otherwise, unless you wanted to be left null. Then negative numbers are turned into 0.

Data Flow for Sales



Data Flow Task: Sales Project 1

Control Flow



Brief Summary

I just went through and fixed the issues where I found them, it is to what I would fix it to, rather than exactly to the instructions. I'm saying this because for the date, I just have any nulls showing that date, not exactly to the format. I used DT_DBDATE for it. I started with Customers data and you can see in the data flow there are a lot of Derive Columns. That is due to me not realizing I can't have two of one table in one derive, but I can have as many columns as long as I don't repeat them. So that's why the Data Flow for Customers is so much bigger, not only from the fact that there were more to fix. But due to the fact I realize I could hit multiple columns in one derive.