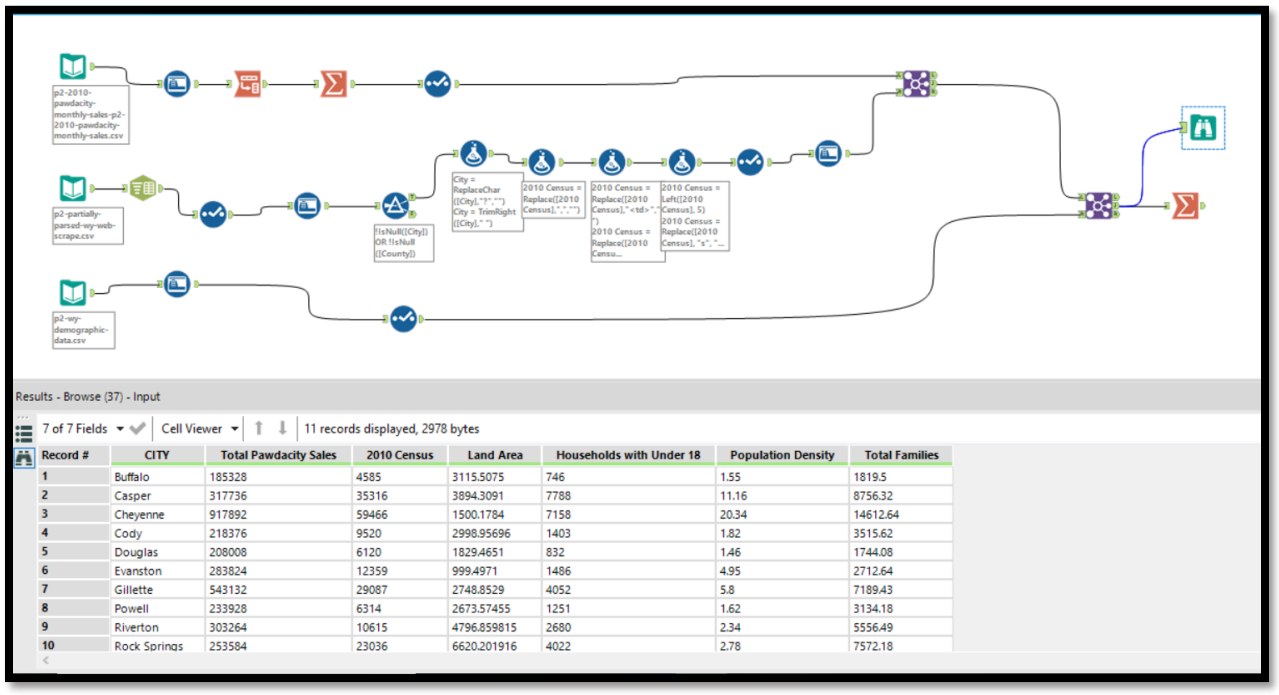
Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. They're looking to expand and open their 14th store. As a Business Analyst, I'm required to predict where this 14th store must be opened based on predicted yearly sales.
The data provided is:

a) Monthly sales data for all the Pawdacity stores for year 2010.
b) NAICS data on the most current sales of all the competitor stores.
c) Parsed file to get population numbers.
d) Demographic data for each city: Households with individuals under 18, Land Area, Population Density, and Total Families.

Step 2: Building the Training Set



Record #	CITY	Total Pawdacity Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
2	Casper	317736	35316	3894.3091	7788	11.16	8756.32
3	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
4	Cody	218376	9520	2998.95696	1403	1.82	3515.62
5	Douglas	208008	6120	1829.4651	832	1.46	1744.08
6	Evanston	283824	12359	999.4971	1486	4.95	2712.64
7	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
8	Powell	233928	6314	2673.57455	1251	1.62	3134.18
9	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
10	Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18

Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,097*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.7*
Total Families	*62,653*	*5,696*

Step 3: Dealing with Outliers

Yes, there are three cities having fields which are outlier.
I've marked red all the outliers (in the picture below).

As we can observe, 'Rock Springs' has only one outlier field which is 'Land Area'. Land Area of one city can be more than other cities with great difference, so this can be ignored.
'Gillette' has 'Total Sales' which is the field we are trying to predict as an outlier. This can pose as a problem to our business decision.
'Cheyenne' has four out of six fields as outliers. These numbers make sense as high population in a small land area leads to high population density and number of families. But since there are so many fields that are outlier for this city, it makes sense to remove this from the dataset.
If the dataset would have been bigger than just 11 rows, I would have removed 'Cheyenne' and imputed 'Gillette'.

So, my decision is to remove **'Cheyenne'** from the dataset.

	A	B	C	D	E	F	G
1	City	Total Sales	2010 Census Population	Land Area	Household under 18	Population Density	Total Families
2	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
3	Casper	317736	35316	3894.3091	7788	11.16	8756.32
4	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
5	Cody	218376	9520	2998.95696	1403	1.82	3515.62
6	Douglas	208008	6120	1829.4651	832	1.46	1744.08
7	Evanston	283824	12359	999.4971	1486	4.95	2712.64
8	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
9	Powell	233928	6314	2673.57455	1251	1.62	3134.18
10	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
11	Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
12	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71
13							
14							
15	Q1	226152	7917	1861.721074	1327	1.72	2923.41
16	Q3	312984	26061.5	3504.9083	4037	7.39	7380.805
17	IQR	86832	18144.5	1643.187226	2710	5.67	4457.395
18	UF	443232	53278.25	5969.689139	8102	15.895	14066.8975
19	LF	95904	-19299.75	-603.059765	-2738	-6.785	-3762.6825