

## Project 2.1: Data Cleanup

### Step 1: Business and Data Understanding

1. What decisions needs to be made?

We want to predicted yearly sales to decide in which city the Pawdacity company can open a new store.

2. What data is needed to inform those decisions?

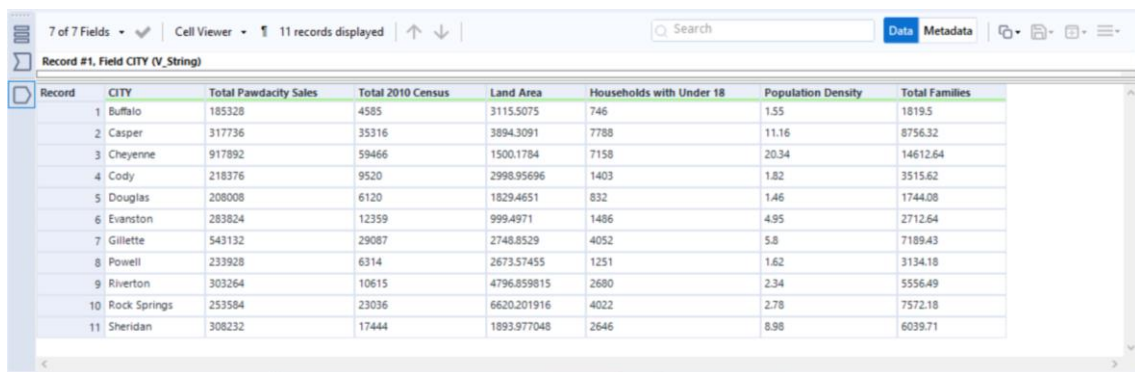
To make the dissection in the above question we need to know:

- 1- City
- 2- Census population of the city
- 3- Households with under 18
- 4- Land area
- 5- Population density
- 6- Total Families

And the target will be total sales.

### Step 2: Building the Training Set

After cleaning and blending different data sets to fill this table here the final training set with 11 records and 7 fields:



Record	CITY	Total Pawdacity Sales	Total 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	283624	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
11	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

Here the sum of each column:

The screenshot shows a data viewer interface with a table containing 7 columns. The first column is labeled 'Record' and contains the value '1'. The other columns are labeled with their respective sum values: 'Sum\_Total Pawdacity Sales' (3773304), 'Sum\_Total 2010 Census' (213862), 'Sum\_Land Area' (33071.380389), 'Sum\_Households with Under 18' (34064), 'Sum\_Population Density' (62.8), and 'Sum\_Total Families' (62652.79).

Record	Sum_Total Pawdacity Sales	Sum_Total 2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families
1	3773304	213862	33071.380389	34064	62.8	62652.79

And the average of each column:

The screenshot shows a data viewer interface with a table containing 7 columns. The first column is labeled 'Record' and contains the value '1'. The other columns are labeled with their respective average values: 'Avg\_Total Pawdacity Sales' (343027.636364), 'Avg\_Total 2010 Census' (19442), 'Avg\_Land Area' (3006.489126), 'Avg\_Households with Under 18' (3096.727273), 'Avg\_Population Density' (5.709091), and 'Avg\_Total Families' (5695.708182).

Record	Avg_Total Pawdacity Sales	Avg_Total 2010 Census	Avg_Land Area	Avg_Households with Under 18	Avg_Population Density	Avg_Total Families
1	343027.636364	19442	3006.489126	3096.727273	5.709091	5695.708182

Column	Sum	Average
Census Population	213,862	19442
Total Pawdacity Sales	3,773,304	343027.63
Households with Under 18	34,064	3096.72
Land Area	33,071	3006.48
Population Density	63	5.70
Total Families	62,653	5695.70

### Step 3: Dealing with Outliers

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? explain your reasoning.

To find the outliers I calculated the upper fence and the lower fence for each column as shown in the table below:

	Total Pawdacity Sales	Total 2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Q1	226152	7917	1861.721	1327	1.72	2923.41
Q3	312984	26061.5	3504.908	4037	7.39	7380.805
Interquartile Range	86832	18144.5	1643.187	2710	5.67	4457.395
upper fence	443232	53278.25	5969.689	8102	15.895	14066.8975
lower fence	95904	-19299.75	-603.06	-2738	-6.785	-3762.6825

So, the outliers were in three cities: Cheyenne, Gillette, And Rock Springs as shown in the data set below:

1	CITY	Total Pawdacity Sales	Total 2010 Census	Land Area	Households with 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

Do you to the data set countians only 11 cities so we need to remove only one city which is Cheyenne because it's the only city has four outliers, also the other outliers for the other cities are close to upper fence.