# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

#### **Key Decisions:**

Answer these questions

- What decisions needs to be made?
  We should find out which city is suitable for opening a new Pawdacity's newest pet store based on yearly sales.
- 2. What data is needed to inform those decisions? Yearly sales for each city, population, number of families and whether a certain age group buy the most from the pet store. Also, how much competitors sell in other cities.

# Step 2: Building the Training Set

7 of 7 Fields 🕶 💜 Cell Viewer 💌 11 records displayed   🔨 🔱						Q Search	
ord	City	2010 Census Population	Total Sales	Households with Under 18	Land Area	Population Density	Total Families
	1 Buffalo	4585	185328	746	3115.508	1.55	1819.5
	2 Casper	35316	317736	7788	3894.309	11.16	8756.32
	3 Cheyenne	59466	917892	7158	1500.178	20.34	14612.64
	4 Cody	9520	218376	1403	2998.957	1.82	3515.62
	5 Douglas	6120	208008	832	1829.465	1.46	1744.08
	6 Evanston	12359	283824	1486	999,4971	4.95	2712.64
	7 Gillette	29087	543132	4052	2748.853	5.8	7189.43
	8 Powell	6314	233928	1251	2673.574	1.62	3134.18
	9 Riverton	10615	303264	2680	4796.86	2.34	5556.49
1	0 Rock Springs	23036	253584	4022	6620.202	2.78	7572.18
1	1 Sheridan	17444	308232	2646	1893.977	8.98	6039.71

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

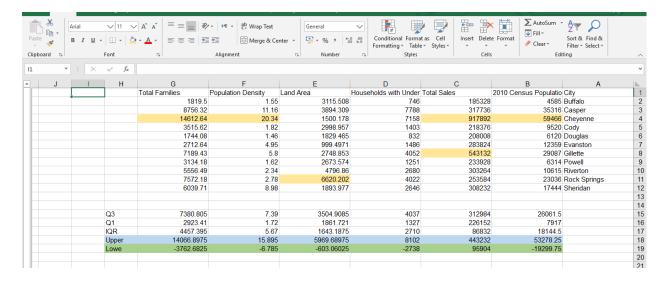
In addition, provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

Column	Sum	Average
Census Population	213,862	19442
Total Pawdacity Sales	3,773,304	343028
Households with Under 18	34,064	3097
Land Area	33,071	3006
Population Density	63	6
Total Families	62,653	5696

## Step 3: Dealing with Outliers

Answer these questions

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.



- 1- Since Cheyenne scored higher than upper fence in census, households with under 18, population density and total families, the given data is most likely legitimate considering the larger number of population density, census and total families.
- 2- Rock Springs have a bigger land area; however, its other values are reasonable.
- 3- Regarding Gillette sales, they seem to be up normal, but there could be other factors that contributed in generating this amount of sales.

In conclusion, Cheyenne has the most outliers therefore we can remove it from dataset.