# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

*The key decision that we need to make in this project is to choose a location for Pawdacity's new store. Particularly, we need to decide in which city to open the store. The decision should be based on the predicted yearly sales of the city.*
*For this reason, we need data about the following:*

- *monthly sales of all Pawdacity stores in different cities,*
- *sales data about competitor stores*
- *demographic data of different cities, including: population number, density, number of families, etc.*

## Step 2: Building the Training Set

| 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

*After using IQR method for each attribute of the cities we find out that from 11 cities Cheyenne and Gillette have outliers in some of them. In fact, Cheyenne has outliers in four of them (Sales, 2010 Census, Population Density and Total Families), while Gillette has only in Sales, moreover, its value is closer to upper fence than the outlier value of Cheyenne.*

| City | Sales | 2010 Census | Land Area | Households with Under 18 | Population Density | Total Families |
|---|---|---|---|---|---|---|
| Buffalo | 185328 | 4585 | 3116 | 746 | 2 | 1820 |
| Casper | 317736 | 35316 | 3894 | 7788 | 11 | 8756 |
| Cheyenne | 917892 | 59466 | 1500 | 7158 | 20 | 14613 |
| Cody | 218376 | 9520 | 2999 | 1403 | 2 | 3516 |
| Douglas | 208008 | 6120 | 1829 | 832 | 1 | 1744 |
| Evanston | 283824 | 12359 | 999 | 1486 | 5 | 2713 |
| Gillette | 543132 | 29087 | 2749 | 4052 | 6 | 7189 |
| Powell | 233928 | 6314 | 2674 | 1251 | 2 | 3134 |
| Riverton | 303264 | 10615 | 4797 | 2680 | 2 | 5556 |
| Rock Springs | 253584 | 23036 | 6620 | 4022 | 3 | 7572 |
| Sheridan | 308232 | 17444 | 1894 | 2646 | 9 | 6040 |
| **1st quartile** | **226152** | **7917** | **1861.5** | **1327** | **2** | **2923.5** |
| **3rd quartile** | **312984** | **26061.5** | **3505** | **4037** | **7.5** | **7380.5** |
| **IQR** | **86832** | **18144.5** | **1643.5** | **2710** | **5.5** | **4457** |
| **Upper fence** | **443232** | **53278.25** | **5970.25** | **8102** | **15.75** | **14066** |
| **Lower fence** | **95904** | **-19299.75** | **-603.75** | **-2738** | **-6.25** | **-3762** |

*As Cheyenne has outliers in 4 out of 6 attributes particularly values that are above the upper fence, it means that we can remove this city to have more correct results in our analysis.*