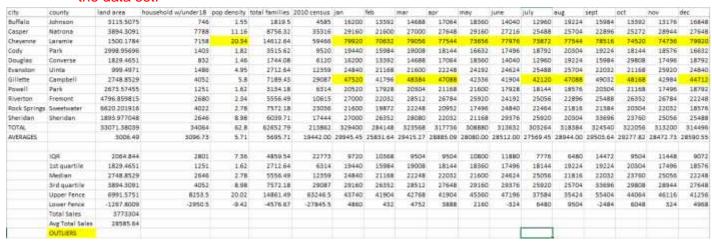
Project 2.1: Data Cleanup

By: Anthony Nguyen

Step 1: Business and Data Understanding

Key Decisions:

- 1. What decisions needs to be made?
 - The decision to select a suitable city for Pawdacity to open a new 14th store.
- 2. What data is needed to inform those decisions?
 - The following data is needed from the current 11 cities in which Pawdacity currently holds locations:
 - Land Area
 - Households With Under 18
 - Population Density
 - Total Families
 - o 2010 Census
 - Total Pawdacity Sales
 - From this data, we determine which city holds outliers in which we can remove from the data set.



Step 2: Building the Training Set

Column	Sum	Average
Census Population	213,862	19,442.00
Total Pawdacity Sales	3,773,304	28,585.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.49
Population Density	63	5.71
Total Families	62,653	5,695.71

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? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

- Yes, there are outliers in the training set = Cheyenne + Gillette
- The outlier chosen to remove = Cheyenne.
- Reasoning for removing Cheyenne is that nearly all of the variables are almost double or nearly meet the Upper Fence compared to the 2nd city that has variables outside the upper fence, Gillette.
 - o Gillette is the 2nd city which has outliers in 1 variable, sales. 7/12 months of sales are outliers.
 - Cheyenne has outliers in 2 variables, population density + sales. 12/12 months of sales are outliers.