Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

In this Business problem, we need to recommend the most profitable city for a leading pet store, "Pawdacity", to open their 14th outlet by predicting their yearly sales in that city.

The data needed must be the monthly sales of all Pawdacity stores for the whole year. The data of number of sales in cities by competitor stores and the demographics of those cities is also necessary to make sure that we obtain a strong relation with the targeted variable. The demographics of those cities must include Households with Under 18, Land Area, Population Density, Population Numbers and Total Families

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

After applying the box and whiskers method on the training set obtained earlier, the data of the 11 cities is compared to the upper fence and lower fence. After analyzing all the predictor variables, two cities appear to be in the outlier category; Cheyenne and Gillette.

The city that should be removed is Gillette. Although Cheyenne is the outlier in more variables, the city has $917,892 sales with a population density of 20.34 while Gillette has $543,132 sales with a population density of 5.8. Both are outliers as the upper fence for sales is $466,776 but Gillette appear to have less logical sales compared to its population size.