Project 2.1: Data Cleanup

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   The objective of the analysis is to provide my manager with the right city to expand by opening the 14th branch of the pet store.

2. What data is needed to inform those decisions?
   Since the manager asked me to inform him the best place based on the historical yearly sales among the cities, the important data is the historical yearly sales and variables that affect the sales; therefore, total population, the population based on ages and area of the store are all increase the model accuracy to predict the sales and make the decision.

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

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.

After using IQR method to find the outliers cities the method results in two outliers which are Cheyenne and Gillette. However, I decided to remove Cheyenne only from the analysis while keeping Gillette for two reasons; Cheyenne is dramatically different than other values; moreover, Cheyenne is an outlier for two variables which are “Total sales and Population Density” and its value in the most of the variables (Except Land Area) is very high and it is near from the upper fence value so I preferred remove this city instead of impute it. On the other hand, despite of Gillette is an outlier but it’s not significantly different as much as Cheyenne. Also, Gillette is an outlier only for one variable and its value is reasonable in the other variables.