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

1. What decisions needs to be made?
   Answer: the business decision that needs to be made here to determine the city in Wyoming to open Paw city's newest branch for the most sales/revenue.
2. What data is needed to inform those decisions?
   Answer: Data on the present location of the present stores and sales, data on prospective store locations and information on the factors that affect sales like the population, competitors, pet population.

Step 2: Building the Training Set

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*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
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

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.
Answer: There are two outlier cities in the dataset; Gillette for total paw city sales field and Cheyenne for total paw city sales, population density and total families fields. The outlier that I have chosen to drop is the City Gillette. On closer observation of both outliers, for Cheyenne, the data about the total sales is consistent with its other information like population density, and

total families which are also outliers. The figures for population density and total families also explains why the total sales is so high. While for Gillette the size of the sales there does not fit with the rest of the information about the city and other cities with similar data like it.