Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state .
for expanding they decided to open thier new 14th store . we will recommend the city for Pawdacity's newest store by predict the yearly sale based on other stores sales and the geographic data

Key Decisions:

Answer these questions

- What decisions needs to be made?
  which city is recommended to open the 14th store in by predicting the yearly sale
- What data is needed to inform those decisions?
  - The sales record for all Pawdacity stores for 2010.

  population numbers(2010 Census Population)
  - demographic data for each city and county in Wyoming(Households with Under 18,Land Area,Population Density,Total Families)

  to create the training dataset:

  City
  2010 Census Population
  Total Pawdacity Sales
  Households with Under 18
  Land Area
  Population Density
  Total Families

- the most current sales of all competitor stores

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*	343,028
Households with Under 18	*34,064*	3,097
Land Area	*33,071*	3,006
Population Density	*63*	6

Total Families	*62,653*	5,696

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.

I took two ways for determine which cities considered as an outlier box and whisker in the scatter plot tool ,then to be more accurate i used IQR method
and found in *Total Families and Population Density* variables that Cheyenne city is an outlier
in *Total Pawdacity Sales* variable Cheyenne and Gilletee cities are outlier
so decided to remove Cheyenne city