Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   Pawdacity, a leading pet store chain in Wyoming, needs recommendation on where to open its 14th store.

2. What data is needed to inform those decisions?

   To properly build a model predictor variables City, Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density and Total Families are suggested to be used.

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

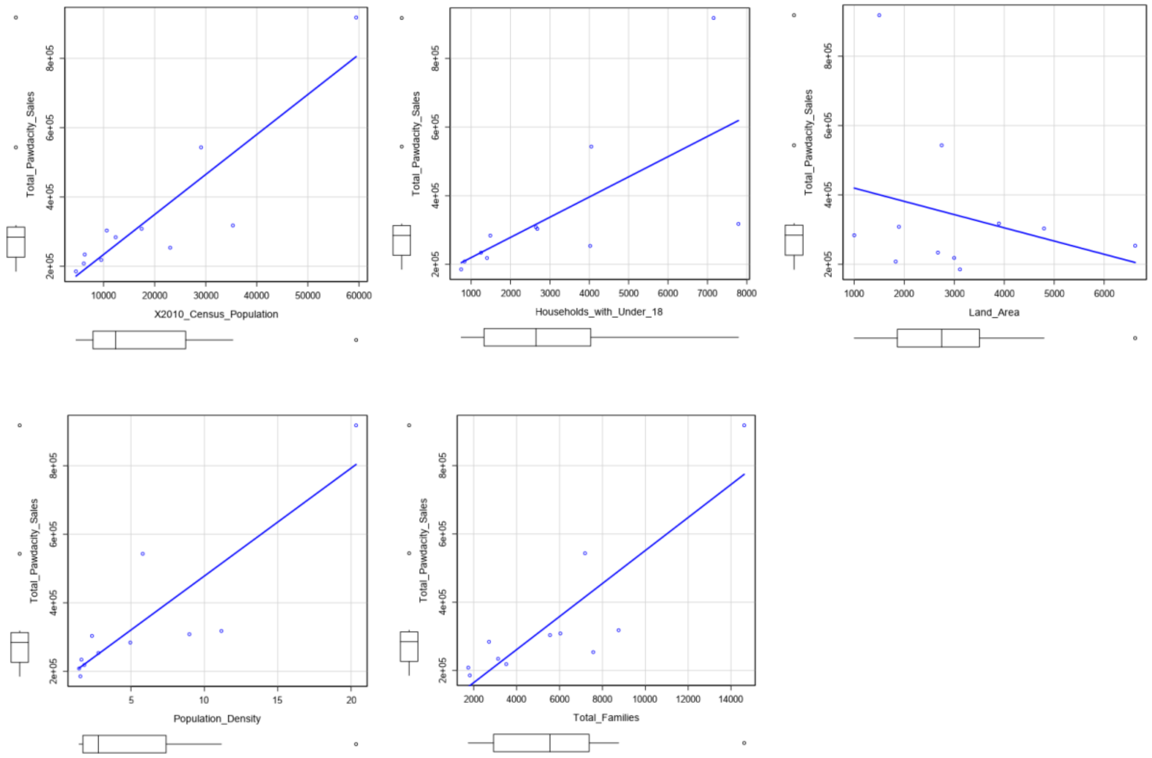
Answer these questions

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.

By checking outlier city in their box and whisker plot for each numeric data

1. **Total Pawdacity Sales:** Gillette city and Cheyenne
2. **2010 Census Population:** Cheyenne.
3. **Households with under 18 years:** none
4. **Land Area:** Rock Springs
5. **Population Density:** Cheyenne
6. **Total Families:** Cheyenne

Refer the charts below



Since Cheyenne is repeatedly appeared as an outlier.It is best suggested to be removed.