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?

   The business decision that needs to be made is whether Pawdacity, a leading pet store chain, should expand and open a 14th store. The decision point is to choose the city in which to expand and start the newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

   To inform these decisions, we need to predict the yearly sales for the cities using a linear regression model. The data needed for this includes the total yearly sales in the current cities in which Pawdacity operates, the 2010 population data for these cities, number of households with individuals under 18, the total number of families as well as the land area and population density in these cities. From these, the predictor variables will be chosen and used to build the regression model to inform the decision-making process.

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

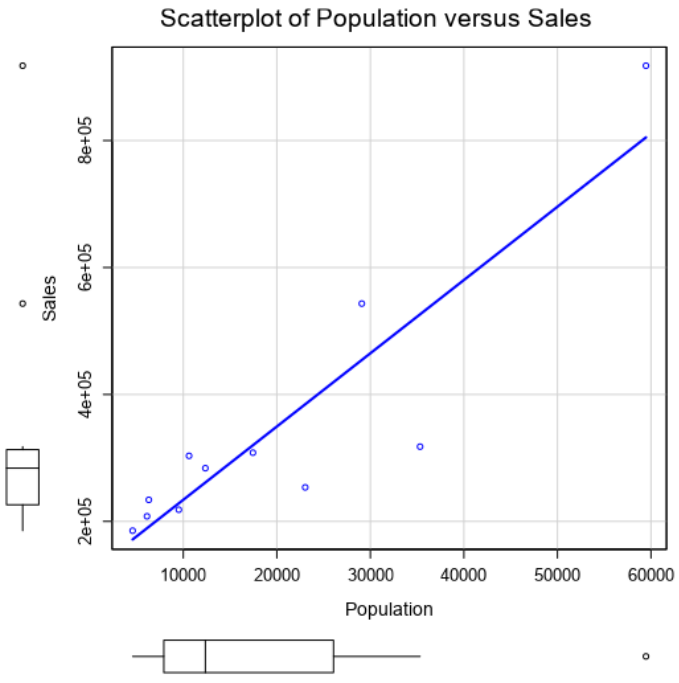
Answer these questions

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.

There are two cities are outliers in the training set:

- **Gillette** for its Pawdacity sale value of 543,132 that is slightly higher than the Upper Fence of 466,776. This outlier is not very significant and based on the fitted line, the outlier is almost in line with the relationship, so we'd leave it in.

- **Cheyenne** for its Pawdacity sales value of 917,892 that is significantly higher than the Upper Fence of 466,776 – almost double the value. Also, its population density (20.34) is marginally above the Upper fence of 20.02. It must also be noted that its 2010 population (59,466) comes close to the Upper fence of 63,246.5. The population density outlier is only marginal and can be ignored but the sales value is a big outlier and extremely beyond the Upper fence value. This skews the model dramatically and will likely hurt the model's ability to make predictions. Looking at the scatter plot indicates this skew. The top right point indicates Cheyenne.



Since the data is far outside the fence and multiple data points are causing the skew, imputing the value doesn't make sense. Therefore, **it has been chosen to remove this data point (city of Cheyenne) from the training set.**