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:

1. What decisions needs to be made?
    - Pawdacity wants to expand its business by opening a 14th store. A decision needs to be made on which city to open the new store based on recommendation.

2. What data is needed to inform those decisions?
    - The data needed to inform the decision include;
        - Yearly sales data for the existing 13 Pawdacity stores
        - Population census of the state of Wyoming
        - Demographic data

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

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 carefully inspecting the training data, I observed that four out of the seven columns contain at least one outlier value.
    I. In the **Census Population** column, the city of **Cheyenne** has a population value of 59,466 which falls in the outlier zone after calculating using the formula for determining outliers(*Q3 + 1.5* IQR). From my calculation, any Population Census value above 53,278.25 is considered an outlier.
    II. In the **Pawdacity Sales** column, there appears to be two outlier values after calculating using the outlier determining formula. Based on my calculation, any Sales value above 443,232 is considered an outlier. The two outlier cities in the Pawdacity Sales column are: **Cheyenne** with a sales value of 917,892 and **Gillette** with a sales value of 543,132.
    III. In the **Land Area** column, one outlier is present, which is the city of **Rock Springs** with a Land Area value of 6620.20. Applying the outlier determining formula, any value above 5,969.7 is considered an outlier.
    IV. In the **Population Density** column, the city of **Cheyenne** with a Population Density value of 20.34 happens to be an outlier. Its value goes beyond the outlier value of 15.9 as calculated using the outlier determining formula.

- From the observation above, the city of **Cheyenne** has a consistent outlier values as compared to the other three outliers.
- I have decided to exclude the city of Cheyenne from my analysis. This is because the data will most likely introduce some ambiguities into the model if considered.
- Because there is a limited amount of data to work with and excluding any further data will greatly affect my model, I will proceed to include the other three outliers in my model and explore the best possible way to achieve the best result using them.