Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions needs to be made?

I have to perform an analysis to recommend the city for Pawdacity's newest store. This I should do based on predicted yearly sales. Knowing that Pawdacity is a leading pet store chain in Wyoming with 13 stores and wish to expand and open a 14th store.

2. What data is needed to inform those decisions?

These are the data I need to inform such decisions:

- the monthly sales data for all of the Pawdacity stores
- current sales of all competitor stores
- the population numbers for each city and
- the Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming.

Using the data source above, I need to create a dataset with the following columns to inform the decisions that we want to make: City, 2010 Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, Total Families.

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*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

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.

To calculate the upper fence and the lower fence, I used the following steps:
I calculated

1. 1st quartile Q1 and 3rd quartile Q3 of the dataset using the Excel function QUARTILE.INC or QUARTILE.EXC
2. The Interquartile Range: IQR = Q3 - Q1
3. For the upper fence: I added Q3 + 1.5 IQR
4. For the lower fence, I subtract Q1 - 1.5 IQR

So therefore, the values above the Upper Fence and below the Lower Fence are the outliers and they are Cheyenne, Gillette, and Rock Springs Cities.

Analyzing the data for the different cities/outliers to check which one to remove;

Cheyenne outlies in Total sales, Population Density, total families and total population, this could be because it can be a big city to have the above outliers. I will ignore this

Rock Springs outlies in Land Area which I could assume that some cities have large land areas.

However, **Gillette** outlies only in total sales with all other parameters in the interquartile range. This I will remove because it seems abnormal to have high sales in within normal range of the other parameters.