Creating an Analytical Dataset

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**

   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state and now, would like to expand and open a 14th store, so we have to predict the city for Pawdacity's new store by analyzing the previous year sales of each city.

2. **What data is needed to inform those decisions?**

   To inform those decisions, we need these data:
   1) monthly sales for all the Pawdacity stores
   2) most current sales of all competitor stores
   3) population records of each city
   4) Demographic data (Households with individuals under 18, Land Area, Population density and Total Families) for each city and county in the state of Wyoming

Step 2: Building the Training Set

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

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 identify the outlier, I first calculate the upper fence and the lower fence for each metric. Here is what I did:

1. Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. I use the Excel function QUARTILE.INC.
2. Calculate the Interquartile Range: IQR = Q3 – Q1
3. Add 1.5 * IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5 * IQR
4. Subtract 1.5 * IQR to Q1 to get the lower fence: Lower Fence = Q1 – 1.5 * IQR
5. Compare each value with upper fence and lower fence. The ones that are not in the range are outliers.

There are 3 cities that are outliers I get from the process I mentioned above: **Cheyenne**, **Gillette** and **Rock Springs**.