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 need to be made?
   Pawdacity (a pet store) would like to expand and open its 14th store in Wyoming. The decision that needs to be made is the choice of the city where the new store will be opened. The decision will be based on a predicted analysis.

2. What data is needed to inform those decisions?

The needed data comes from 3 different datasets :

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
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.

The data needed to be cleaned, formatted then gathered into one file by joining it.

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.

After calculating the interquartile, the lower and upper fence. I found out that there are 3 cities with outlying values in the dataset.
The cities are Cheyenne, Gillette, Rock Springs.
The outlier city that I decided to remove was Cheyenne. Because its values were higher in 4 different fields while the two other cities each had only one outlying value.
Adding to that the outlying values for Cheyenne were far greater than the upper fence.
The values are shown in the table below:

CITY	Total Pawdacity Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
Upper Fence	443232	53278.25	5969.689139	8102	15.895	14066.8975