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?
   Which city will be the best place to open the Pawdacity's 14th store?

2. What data is needed to inform those decisions?
   - Data on the status of sales from each Pawacity's stores
     - average sale data of last year
   - Data on population of each count in Wyoming
   - Data on households with under 18 and total families in Wyoming counties
   - Land area of each city
   - Population density of each city
   - Sale data of all competitors store

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*	*19,442.00*
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.

Two cities have outliers.

- Cheyenne (917,982 in Total Sales, and 20.34 in Population Density)
- Gillette (543132 in Total Sales)

The mean of Total Sale amount is 343,027.64 and standard deviation is 203,601.17. Both values are away from mean value (z-score 0.98 and z-score 2.82). So I decide not to impute with mean value.

When comparing two cities,

- the increased sale amount in Cheyenne is associated with increased Census Population (z=2.52), increased HHU18 (z=1.74), increased Population Density (z=2.62) and increased Total Families (z=2.45).
- For Gillette, the Sale Amount is increased while other variables are not very much increased. (Census Population (z=0.61), HHU18 (z=0.41), Population Density (z=0.02), Total Families (z=0.41).

The outlier in Gillette skew high only in Total Sale, while outlier in Cheyenne skew high not only in Total Sale but in other variable, as well.

So I decided to remove the city “Gillette”