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?
   To recommend the city for Pawdacity's newest store, based on predicted yearly sales.

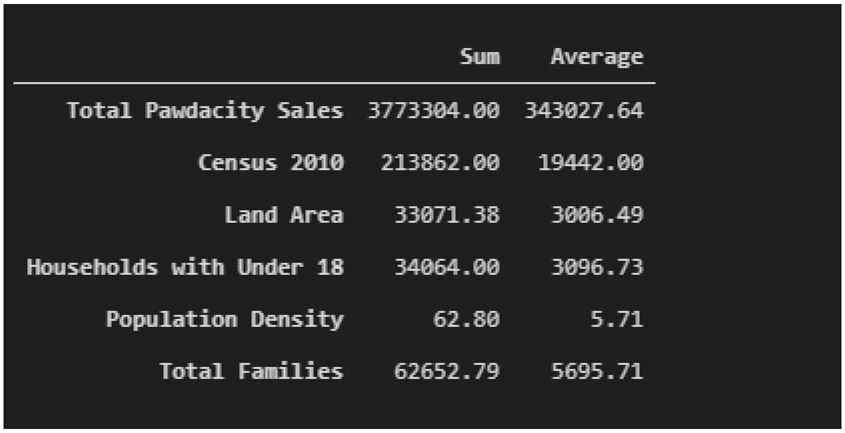
2. What data is needed to inform those decisions?

   - ✓ Pawdacity Monthly Sales
   - ✓ Population Data
   - ✓ Demographic Data
   - ✓ Competitor Sales

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



```
                               Sum    Average
-----------------------------------------------
   Total Pawdacity Sales  3773304.00  343027.64
             Census 2010   213862.00   19442.00
               Land Area    33071.38    3006.49
Households with Under 18    34064.00    3096.73
      Population Density       62.80       5.71
          Total Families    62652.79    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.

Yes. **Cheyenne** is the outlier city because the **Total Pawdacity Sales**, **2010 Census Population**, **Population Density**, and **Total Families** have extreme value which has higher chance to never repeat in other cities. Using this city data will add bias to our model.