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?

   Since , 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 .

: Awesome: Good job identifying the key decision to be made.

2. What data is needed to inform those decisions?

   To inform those decisions , we need to have data of monthly sales for all the Pawdacity stores . We need the data on the most current sales of all competitor stores . We need to have the data which contains the population records of each city . We need 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 .

: Awesome: These data should be adequate to carry out the analysis.

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*

: Awesome: The correct averages have been reported for each column.

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.

There are 3 cities that are outlier in the training set which are : Cheyenne , Gillette and Rock Springs .
I made scatter-plots of CITY with all other variables of all cities . Then, I made scatter-plots of Total Pawdacity Sales with other variables for all cities . By analysing the scatter-plots , I came to know there are 3 cities which have outliers . I ignore the city Cheyenne because it can be a big city (the city Cheyenne outlies in total sales , total population , population density and total families) . I ignore the city Rock Springs because it can happen that some city has larger land area (city Rock Springs outlies in land area) . So I decide to remove the city Gillette which outlies in total sales but have all other things in interquartile range and it also seems abnormal to have high sales with all other variables in proper range .

: Awesome: Well done identifying the correct outliers.

: Awesome: We also make the model more robust to the presence of bigger cities in some future analysis, by retaining Cheyenne.

: Awesome: The decision to remove Gillette is absolutely correct. It's population doesn't seem to support the sales that have been reported. If a city has large sales, we would also expect it to have a big population to drive those sales, which isn't the case with Gillette.