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?
   Pawdacity needs to decide where to locate its 14th store based on estimated yearly sales.

2. What data is needed to inform those decisions?
   To do this, it is required to estimate or predict the yearly sales of available locations. This would involve data on previous sales, population and demography of the state as well as pawdacity competitions' sales. In this project, what is required is to prepare a consolidated dataset that will be used to decide the location.

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.
Using the IQR method, it is shown that Rocksprings is an outlier with respect to its land area while Cheyenne and Gillette were outliers in terms of yearly sales. Cheyenne in fact was an

outlier on all variable counts except land area. Thus, it should be removed as it can skew the analysis.
The outliers were identified using the IQR method and were all found to be higher than the upper fence.
Cheyenne was removed because the city appears as an outlier in almost all the column, this can really affect the analysis. Also, since the dataset is small imputation with mean or similar values can be counterproductive.
The other outliers were retained because they were outliers with respect to only one variable. Thus, they may not really affect the analysis. Besides when the model is built, their effect could be quantified and eventually removed or retained depending on how they affect the model.

N.B

To calculate the upper fence and the lower fence, here are the exact steps:

1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset

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 . Values above the Upper Fence and values below the Lower Fence are outliers