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 is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. The manager has asked me to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?
   To be able to predict the suitable city for the new pet store for Pawdacity, the following data are needed for an accurate prediction that will guide the informed decision that needs to be made.
   - Total number of families: The number of families in a given city tells how many potential customers there are in the city.
   - Households with under 18: The number of households that have members below age 18 shows the number of families with the likelihood to buy toys from pet stores.
   - Population density: The density of the city determines how close the pet store is to the potential customers which translates to potential high sales.
   - Population Census: The population of the city determines the number of available to buy toys from the pet store.
   - Land Area: The land area helps calculate the demographics of a city with respect to other cities around.
   - Yearly sales in city: This data helps calculate the expected sales in each city. This data will help make informed decision on where to open the next pet store with regards to revenue generated on sales within a particular city.

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*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.50*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

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.

Outliers:

1. Rock springs city is an outlier in terms of Land area with the highest land area of about 6,629sqkm. Further analysis into land area of Wyoming state shows that Rock Spring's county which is Sweetwater has the highest land area of about 10,395sqkm with a total of six cities. Except for Green River which is the second largest city with about 3,477sqkm, other cities in Sweetwater has a land area less than 150sqkm. From the analysis of Land area in Sweetwater county, it makes sense that although Rocksprings city is an outlier, the data is accurate hence I included the city in the training dataset, more so, Land area does not affect the predicted yearly sales for the new pet store.

2. Cheyenne city is an outlier in terms of yearly sales with highest total yearly sales of about 917,892 among the eleven cities. I included the outlier in the dataset because going through the dataset to find out why there is so much figure in the yearly sales, I noticed that Cheyenne recorded the highest sales among the eleven cities each month in 2010 with an average of about 76,491 yearly sales. This is justified because Cheyenne is the most densely populated city among the eleven cities with a total of 14,613 families. In 2010, Cheyenne has a population census of about 59,466 counts which is the highest population recorded. With the available data, it makes sense that Cheyenne which is the outlier is a correct data hence I included it in the training dataset.

3. Gillette city is an outlier in terms of yearly sales with second highest total yearly sales of about 543,132 among the eleven cities and an average yearly sale of about 45,261.Being the third largest city in terms of population and household under 18, fourth city with the largest number of families and densely populated area, this raises the question why it is the second city with the highest sales year.
   I chose to include the city in the dataset because the monthly sale has been fairly consistent throughout the month of the year with the highest sales made in September with a total sale of about 49,032 and the lowest sales made in February with a total sale of about 41,796. This monthly sales data justifies the high number of yearly sales in 2010.
   To accurately answer the question on why Gillette is the second city with the most yearly sales, data about family income and number of family member under 18 will be needed.