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)

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. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

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. Key Decision is using data provided to recommend the city for Pawdacity's newest store.

2. What data is needed to inform those decisions?
   Below list of data are suggested to be used to better inform the decision.

   - Census Population
   - Total Pawdacity Sales
   - Households with Under 18
   - Land Area
   - Population Density
   - Total Families

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*	*3,096.72*
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.

Record	City	Total_Sales	2010 Census	County	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185,328	4,585	Johnson	3,115.5075	746	1.55	1819.5
2	Casper	317,736	35,316	Natrona	3,894.3091	7,788	11.16	8756.32
3	Cheyenne	917,892	59,466	Laramie	1,500.1784	7,158	20.34	14612.64
4	Cody	218,376	9,520	Park	2,998.95696	1,403	1.82	3515.62
5	Douglas	208,008	6,120	Converse	1,829.4651	832	1.46	1744.08
6	Evanston	283,824	12,359	Uinta	999.4971	1,486	4.95	2712.64
7	Gillette	543,132	29,087	Campbell	2,748.8529	4,052	5.8	7189.43
8	Powell	233,928	6,314	Park	2,673.57455	1,251	1.62	3134.18
9	Riverton	303,264	10,615	Fremont	4,796.859815	2,680	2.34	5556.49
10	Rock Springs	253,584	23,036	Sweetwater	6,620.201916	4,022	2.78	7572.18
11	Sheridan	308,232	17,444	Sheridan	1,893.977048	2,646	8.98	6039.71

Outlier's present are "**Cheyenne**" City for columns Total_Sales, 2010 Census, Land Area & Population Density, "**Rock Springs**" City for column Land Area & "**Gillette**" City for column Total_Sales.

With a deeper look, **Cheyenne** has two stores, its data aggregated prompting it to look like an outlier, but from careful observation, this seemingly looking outlier is not skewing the data as it will be essential in building the predictive model for analyzing store expansion in the city. So, I will recommend keeping this city.

Gillette also having two stores, appears to be an outlier when its data value is aggregated. Using the interquartile range analysis, the Total_Sales for the city appears to be the clearest outlier. Recommending leaving Gillette out of the variables for the predictive model.

Analyzing the interquartile range analysis, I will recommend keeping **Rock Springs**.