Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   A leading pet store chain, Pawdacity, needs recommendation on where to open its 14th store in Wyoming.

2. What data is needed to inform those decisions?

   The data required in order to inform this decision are *city, 2010 census population, Pawdacity sales in other stores, competitor sales, household with under 18, land area, population density* and *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*	*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

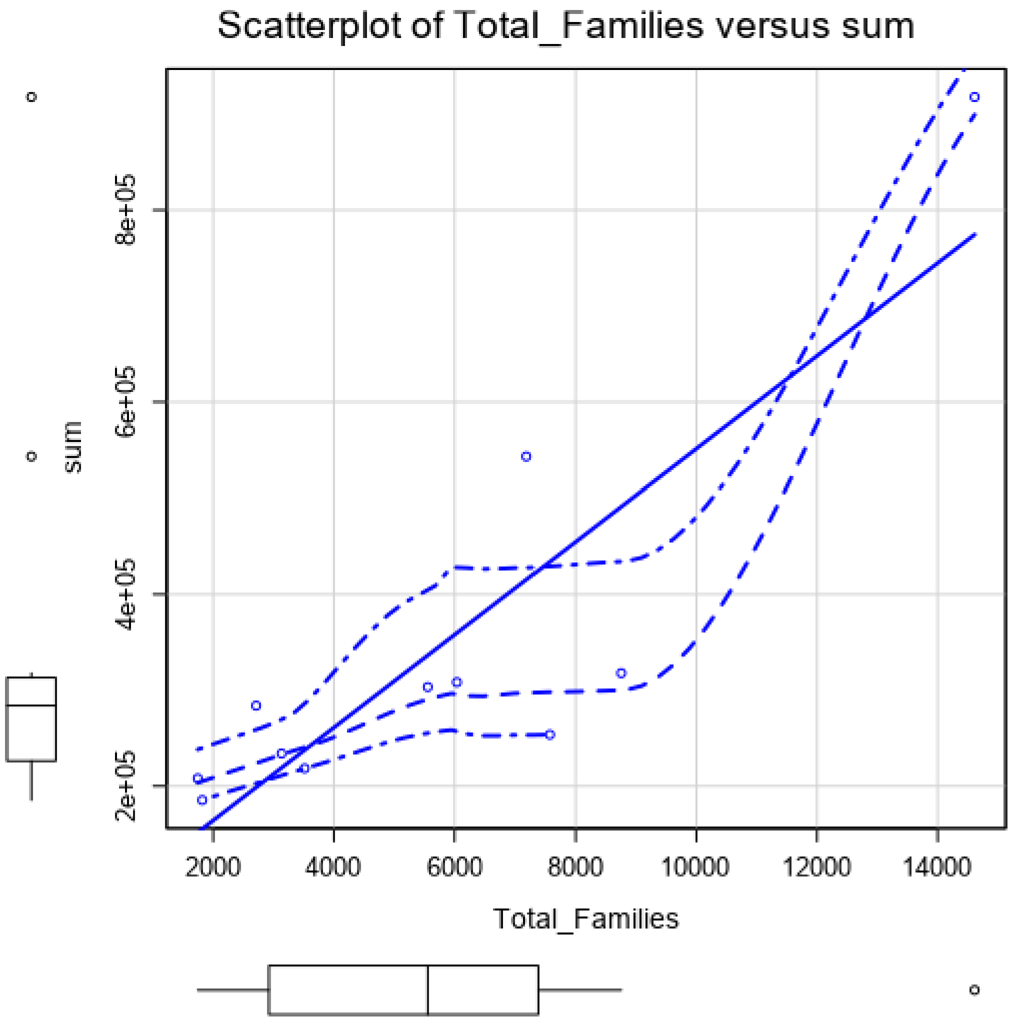
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

Using the IQR method for each attribute, the outliers for each variable are listed below:

Census Population : Cheyenne	
Land Area : Rock Springs	

Population Density : Cheyenne	
Total Families : Cheyenne	
Total Sales : Gillette and Cheyenne	

From the scatterplot for total families vs. total sales, we can see that both Gillette and Cheyenne are outliers. However, Cheyenne also has been outlier in Census Population, Population Density and Total Families. Therefore, Cheyenne should be considered as outlier.


Scatterplot of Total_Families versus sum
sum
2e+05
4e+05
6e+05
8e+05
Total_Families