**Project: Data Cleanup**

## **Step 1: Business and Data Understanding**

1. What decisions need to be made?

The decision to be made is to recommend the city for Pawdacity newest store based on predicated yearly sales. The store would like to expand and open the 14th store. It is important to determine which city to recommend because there are several cities, and we need to choose the city that will provide the best sales for the newest stores.

1. What data is needed to inform those decisions?

We will need data on existing store sales as well as demographic information associated with these existing stores, such as household income, age distribution, education, race, and population density, to see whether there are any demographic variables that can drive our existing store sales.

## **Step 2: Building the Training Set**

|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| Census Population | 213,862 | 19,442 |
| Total Pawdacity Sales | 3,773,304 | 343 027.63 |
| Households with Under 18 | 34,064 | 3 096.72 |
| Land Area | 33,071 | 3 006.45 |
| Population Density | 63 | 5.72 |
| Total Families | 62,653 | 5 695.72 |

## **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?

I used the IQR Method to determine if there are outlier cities for each of the variables.

Yes, there are outliers in the training set. The outliers are Cheyenne, Rock Springs, and Gillette cities. There are outliers because the values are above the upper fence for these variables:

* Census population (Cheyenne)
* Sales by Month (Gillette, Cheyenne)
* Population Density (Cheyenne)
* Land Area (Rock Springs)

However, Cheyenne is the outlier that I choose because it has double the upper fence value concerning the sales by month, and Cheyenne has values above the upper fence in 3 variables which are the sales by month and the census population, and population density. The other outlier, Rock Springs and Gillette have a value above the upper fence for only one variable. Thus, I decide to impute the outlier because we have six variables, and 3 of the variables are above the upper fence. If the number of variables was more than 3, I would have removed the outlier.