Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

1. What decisions needs to be made?

Perform an analysis to recommend the city for Pawdacity’s newest store, based on predicted yearly sales.

1. What data is needed to inform those decisions?

Dataset with the following columns:

City  
2010 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* | *19442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343027.64* |
| *Households with Under 18* | *34,064* | *3096.73* |
| *Land Area* | *33,071* | *3006.45* |
| *Population Density* | *63* | *5.73* |
| *Total Families* | *62,653* | *5695.73* |

## 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.

I used the IQR method to determine if there are outlier cities for each of the variable

Details in excel sheet. Cheynne and Gillette total sales were outliers.

I have removed one outlier, Cheynne, with the filter tool. This is due to that Cheynne had much larger difference with the other cities on sales than Gillette. Cheynne had similar levels of other fields with Gilette with the exception of total families. It is better to remove Cheynne because its irregularities compared to others. Please refer to Charts in excel document.