Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?
    - Based on the predicted yearly sales the management needs to decide a city for opening a new store.

2. What data is needed to inform those decisions?
    - The data required is yearly sales of stores, yearly sales of cities, population density, household under 18, total families and land area. These all variables will help in predicting yearly sales for new cities and where it would be most profitable to open a new store.

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*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

Outliers in total Pawdacity sales:
Cheyenne – 917892
Gillette – 543132

Outliers in 2010 census:
Cheyenne – 59466

Outlier in Land area:
Rock springs - 6620.201916

Outliers in Household:
Casper – 7788
Cheyenne – 7158

Outlier in population density:
Cheyenne – 20.34

Outliers in total families:
Cheyenne – 14612.64

As, we can clearly see city Cheyenne is an outlier in most of the categories. This also does not seem to be an abnormal value as higher sales can be attributed to large number of families and even high number of households with people under 18.
As all the values seem to be in the case of Cheyenne, we will remove the city and do our analysis as then all other variables will be on a similar scale and produce better results. This city will increase the mean of the variables as most of the values are on the higher end.