Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity needs to open the 14th store in Wyoming and in order to know in which city to open the new store I need to combine the three data sets together. The three data sets contains information about the other 13 stores and their monthly sales, information about the areas where these stores are and about the population density.
Before I could join those data sets I had to clean them and change some data type.
After the join I have looked for outliers in the final data set.

Key Decisions:

Pawdacity needs to decide in which City to open a new store. In order to take this decision they need to have information about the other store they own, about the cities this other stores are and about the population of these cities.
They will eventually compare these information with information about other competitors' stores to predict the revenues from the 14th store.

Step 2: Building the Training Set

Column	Sum	Average
Census Population	*213,862*	*19,44*
Total Pawdacity Sales	*3,773,304*	*34,3027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5,7*
Total Families	*62,653*	*5,695.70*

Step 3: Dealing with Outliers

Answer these questions
There are two City with outliers in the training set : Cheyenne and Gilette.
I have decided to remove Cheyenne because it has outliers in "Pawdacity sales volume" and in "Population Density" and they are both way higher then the threshold
The outlier in Gilette is in "Pawdacity sales volume" and I have decided to bring it to the value of 455,112 that is the value of the upper fence.