Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?
   Recommend possible city to open the new store in it.
2. What data is needed to inform those decisions?
   City, Each city's sales (summarize stores' sales in each city), 2010 census population, households with under 18, land area, population density and total number of families.

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*133252*	*12113.82*
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.

There is one outlier which is Cheynne city with following data:

- Total Sales equal to 917892 (exceeds the upper fence = 443232).
- Population Density equal to 20.34 (exceeds the upper fence = 15.90).
- Total Families equal to 14612.64 (exceeds the upper fence = 14066.8975).

Since it's positive outlier that indicates how important its sales and has huge target customers, so we can impute it.