Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Considering the predicted yearly sales as target variable, a decision must be made that recommends the city for Pawdacity's newest store. The relevant data must be formatted and blended from different datasets and outliers are to be dealt accordingly.

2. What data is needed to inform those decisions?

   City or County wise Pawdacity sales, census population, demographic data of Households with under 18, Land area, Population density and Total families. Interquartile range to calculate upper fence and lower fence values so to further determine the outliers affecting the dataset.

Step 2: Building the Training Set

Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.45*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

Step 3: Dealing with Outliers

Cheyenne, Gillette and Rock Springs were identified as outliers in the training set. Interquartile range depicted through box and whisker plots (Scatter plots) was used to determine the cause and effect of outliers.

2010 Census Population: Cheyenne was identified as outlier. The value is above the upper fence and seems to skew the data thereby causing an effect on the relationship with total Pawdacity sales.

Total Pawdacity Sales: Cheyenne and Gillette were identified as outliers. The fitted line associated with Gillette seems to be in line considering its census population.

Households with under 18: No Outliers were identified.

Land Area: Rock Springs was identified as outlier. The data seems to be in line with the trend and not dramatically different.

Population Density: Cheyenne was identified as outlier. The value is above the upper fence and seems to skew the data thereby causing an effect on the relationship with total Pawdacity sales.

Total Families: Cheyenne was identified as outlier. The value is above the upper fence and seems to skew the data.

Considering the variables above, models were built with and without including **Cheyenne**.