Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a pet store chain with 13 stores throughout the state of Wyoming. The manager is considering opening another store in the state. The manager has requested an analysis of historical data to recommend where the new store should be based. In order to complete this analysis, historical sales data from previous sales, along with various city/county metrics is required. This is provided by the manager in the form of historical monthly sales, city/county population numbers and demographic data. The aim of this report is to initially prepare the data to conduct such predictive analysis.

Step 2: Building the Training Set

After cleaning, blending and joining our datasets, a training set was built containing eleven rows of data and six columns, namely Census population, Total Pawdacity sales, Households with under 18, Land area, Population density, Total Families. The following table is a summary of the sum and average values of these columns:

Column	**Sum**	**Average**
Census Population	*213,862*	*19442*
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 were identified by using a threshold defined by 1.5 interquartile ranges above the third quartile or below the first quartile. There were two cities with outlier variables in the dataset. Namely; Gillette and Cheyenne. As Cheyenne has a higher number of outliers, it could be suggested that this city should be removed, however on inspection of the dataset, it appears that the extreme total Pawdacity Sales can be explained by the other

extreme values in this city (e.g. Population density). In contrast, Gillette's outlier value of 'Total Pawdacity Sales' did not coincide with any other features to explain this value.

Rather than removing Gillette from the already sparse dataset, the 'Total Pawdacity Sales' outlier value was regression imputed. To do so, stepwise linear regression was utilized to create a model to predict the outlier value. The model containing three significant predictor values (Land Area, Households with under 18 and Total Families) could explain 95% of the variance in 'Total Pawdacity Sales'.

Total Pawdacity Sales = 172630.55 + Land Area (-42.81) + Households with under 18 (-39.95) + Total Families (72.14)

This model predicted the 'Total Pawdacity Sales' for Gillette as 411720.24. Future reports should explore the use of stochastic regression imputation to avoid overfitting this data point.