Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions needs to be made?

2. What data is needed to inform those decisions?

Pawdacity is a leading pet store chain with 13 stores in the state with the intention to create a 14th store. We want to make a decision to create a 14th store and where it should be located. Below are questions that inform this decision.

Which store has the highest number of yearly sales for Pawdacity in 2010 and what factors contributed to the amount of sales the store had

It is important to determine the store with the highest sales and the city where the store is present. What made it stand out and factors affecting the high number of sales in that store. This can be checked using the monthly sales of 2010.

The demographic of each city in the State, the population density, the Landform and the Population Structure

The demographics, density and the population structure greatly contribute to the amount of sales and what sells in a store. Areas that are densely populated may have more market that lower populated cities. This can be checked using the Demographic Data and the data for population numbers.

The Level of Competition of the different stores in the state and factors affecting competition

Competition could reduce the amount of sales in a particular store. It is important to check the level of competition of a product in a particular city, or ways to stand above the competitors in that city. This can be checked using the NAICS data.

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*
Total Pawdacity Sales	*3,773,304*	343027.64
Households with Under 18	*34,064*	*3097*
Land Area	*33,071*	3006.49
Population Density	*63*	*5.7*
Total Families	*62,653*	*5696*

Step 3: Dealing with Outliers

Answer these questions

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.

The cities with Outliers are Cheyenne and Rock Springs. Since the Cheyenne City has outliers in 3 variables (Census, Total Families and Population Densities) and Rock Springs has just one Outlier (Land Area).

If one city is to be removed, it should be the Cheyenne City since it has 3 outliers, because otherwise, it would affect the shape of the dataset.