

Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Pawdacity would like to open a 14th store in Wyoming. We need to analyze yearly sales data of all competitor stores to determine the city for Pawdacity's newest store.

2. What data is needed to inform those decisions?

We need sales data of all Pawdacity stores, sales data of all competitor stores, population data of Wyoming cities, and demographic data of Wyoming cities.

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	19442.00
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

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.

I would choose to remove the city Cheyenne because it has the most (four) outliers in the demographic and population data. Gillette and Rock Springs only have one outlier each, so I would impute these values.