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?

Recommend the city for newest store based on predicted annual sales.

2. What data is needed to inform those decisions?

2010 census population, total Pawdacity sales, households with under 18, land area, population density and total families for each city.

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*	*3096.73*
Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.27*

Step 3: Dealing with Outliers

Answer these questions

Are there any cities that are outliers in the training set?

Yes.

Which outlier have you chosen to remove or impute?

Cheyenne.

Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

Cheyenne has outliers in 3 feature, i.e. 2010 census population, total Pawdacity sales and total families. Because it is different than the other, I will remove this city as significant outlier.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.