Project 2.1: Data Cleanup

Make a copy of this document. Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

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?
   Ans: In which city should we open our new (14th) pet store based on predicted yearly sales.

2. What data is needed to inform those decisions?
   Ans: Demographic data (City Land Area, Population, Total Families, Population Density, Households with Under 18), Monthly Sales for all Pawdacity Stores, Competitor Sales 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*	*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.

Ans: Yes there are cities with outliers. Cheyenne has outliers for all the predictor variables except Land Area but it also has an outlier for the target variable, Total Paudacity Sales hence it follows the pattern and won't be dropped.
Rock Spring has an outlier for Land Area but this doesn't follow the same pattern with the target variable. Land Area does not have a significant impact on Total Pawdacity Sales so I'll leave this.
Gilette has an outlier for our target variable, Total Pawdacity sales hence I'll drop it.

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.