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)
Pawdacity is a pet store chain and they plan to open a new store. They would like to know which city would be the best option. This option will be determined by perfoming an analysis based predicted yearly sales.

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Pawdacity needs to decide where to open the 14th store.
2. What data is needed to inform those decisions?
   The decision should be based on the following data city, city population, Pawdacity store sales, demographic information of the cities and sales data of competitor business.

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.

The findings for outliers are listed below.

City	Population	Sales	Land Area	Population Density	Total Families
Cheyenne	x	x		x	x
Gillette		x			
Rock Springs			x		

I decided to remove Cheyenne. First of the all, it has significantly bigger population than other cities. The closest one is Casper which has 24,150 population less than Cheyenne. Cheyenne also stand out as an outlier for “total sale”, “population density” and “total families” data. All these flags make sense because it seems Cheyenne is a much bigger city compared to others. Therefore, it is a better decision to compare same scale cities rather than including a large one since there is a limited number of cities in the data.

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.