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 wants to open a new Wyoming store. Based on the information provided from the current 11 stores, predict the sales generated from a new store. Each store is in a different city, so city is the unique identifier for the stores.

Key Decisions:

Answer these questions

1. What decisions needs to be made?
    - The decision is which city to open a new Pawdacity store.

2. What data is needed to inform those decisions?
    - The data needed to inform those decision is the current data from the 11 existing stores. The project details says to use Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, and Total Families.

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.

- I provided scatterplots of the variables, since the scatterplots in Alteryx has box plots. From the box plots, I determined how many outliers there were for each predictor, then found the matching cities that were outliers. For Total Pawdacity Sales, Cheyenne and Gillette were outliers. For 2010 Census, Cheyenne was an outlier. For Land Area, Rock Springs was an outlier. For Total Families, Cheyenne was an outlier. For Population Density, Cheyenne was an outlier. From doing a little background research, Wyoming is the least populous state. Its capitol is Cheyenne. I would expect Cheyenne to be different from almost all other cities in Wyoming. Thus, it makes sense for it to be different than most Wyoming cities, so I will remove it from the dataset.

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.