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?

   Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. They needs a recommendation on which city should the new store open.

2. What data is needed to inform those decisions?

A dataset should created with this following columns :
City
2010 Census Population
Total Pawdacity Sales
Households with Under 18
Land Area
Population Density
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*	*19,442.00*

Total Pawdacity Sales	*3,773,304*	*343,027.60*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
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.

After adding the new columns, cleaning and joining the datasets needed together, this is the final dataset below :

City	Census_Population	household_with_ubder_18	Land_Area	Total_pawdacity_sales	Population_density	Total Families
Buffalo	4585	746	3115.5075	185328	1.55	1819.5
Casper	35316	7788	3894.3091	317736	11.16	8756.32
Cheyenne	59466	7158	1500.1785	917892	20.34	14612.64
Cody	9520	1403	2998.95696	218376	1.82	3515.62
Douglas	6120	832	1829.4651	208008	1.46	1744.08
Evanston	12359	1486	999.4871	283824	4.95	2712.64
Powell	6314	1251	2673.57455	233928	1.62	3134.18
Riverton	10615	2680	4796.859815	303264	2.34	5556.49
Rock springs	23036	4022	6620.201916	253584	2.78	7572.18
Sheridan	17444	2646	1893.977048	308232	8.98	6039.71
Gillette	29087	4052	2749	543132	6	7189.11

There are 2 cities seems to be the OUTLIRES here, Cheyenne and Gillette. Based on the dataset their sales data are higher than the other cities.

Cheyenne : very high population and total sales than other cities. all the data field of Cheyenne are almost higher which make this record an outlier.

Gillette : the demographic records are all within the range, Except a very high total sales. Logically a small city with high population should record a higher number of total sales compared with the other cities.
Gillette city has a potential to skew any models we run. I suggest to remove this city 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.