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?

Which city shall the company open its 14th pet store in? This decision is made based on the predicted pawdacity yearly sales of any chosen city.

2. What data is needed to inform those decisions?

At the city level: total pawdacity sales of all pawdacity stores in the city, 2010 total population, land area, number of households with any child under 18, population density, total number of 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*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
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*	*5,695.71*

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 examining scatterplots for all the numerical fields two potential outliers are found

	Cheyenne	Gillette
Total_Pawdacity_Sales	917,892	543,132
2010 Population Census	59,466	29,087
Land Area	1,500.18	2,748.85
Households with under 18	7,158	4,052
Population Density	20.34	5.8
Total Families	14,612.64	7,189.43

The total pawdacity sales in these two cities are much larger than other cities in the dataset. However, other fields of Gillette seem normal, so Gillette shall not be regarded as an outlier. On contrary, this is not true for Cheyenne which has too large values for all but the land area field. This means the city characteristic of Cheyenne is fundamentally different from other cities in the dataset. From these field values, we can see Cheyenne is a relatively large city and its inclusion in the linear regression model significantly changes the result. So in order to understand better other smaller cities, we shall remove Cheyenne from the linear regression model.

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.