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?
   In which city to open the new store is the decision to be made.

2. What data is needed to inform those decisions?
   We have several predictor variables. We need to base on those variables make a model and pick the right predictor variables for the model. And then use the model to predict the city with the potential good sales.

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*	*19442.00*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3006.45*
Land Area	*33,071*	*3096.73*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5695.73*

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.

Based upon the scatter spots and excel calculation, we have the following observation.
For the linear regression relationship between Census population and total sales, the data of Cheyene is an outlier based upon the scatter spot but not an outlier based upon calculation.
For the linear regression relationship between Land Area and total sales, the Rock-Springs data is an outlier based upon the scatter spot but not an outlier based upon calculation.
For the linear regression relationship between Households with Under 18 and the total sales, there is no outlier.
For the linear regression relationship between Population Density and the total sales, Cheyene is the outlier under both the scatter spot and calculation exam.
For the linear regression relationship between Total Families and the total sales, Cheyene is the outlier under both the scatter spot and calculation exam.
We should remove the Cheyene data. I do not think the data was wrong. It may just because Cheyene is extremely dense-populated with much more families and population than usual cities. I do not think we need to impute it since it is a small data pool. We could just use all other actual data to predict and remove this extreme case.

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.