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.

: Awesome: Correct! This is indeed the main business decision to be made.

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

: Awesome: Good work identifying this. This data should be good enough for part 1 of the analysis.

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*

: Awesome: well done! All the sum & averages are perfectly correct!

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.

There is one outlier in the training set: Cheyenne and we shall remove it. Data of this city contains several entries that are outliers based on IQR. These entries are 59466 (2010 Census), 917892(total sales), 7158(under 18), 20.34(population density), 14612.64(total family). All these values are too large compared to values of all other cities.

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.

: Required:
- The decision to remove Cheyenne is acceptable.
- The reasoning is also on the right track.
- But it can be expanded further
- You must reason with outliers based on their characteristics as a city
- For example :"Cheyenne seems to be a big city , in midst of a dataset that contains small and medium sized cities. It has multiple outlier fields and even its other field values , are unlike the other cities in the dataset. Therefore, it can possibly skew our predictor model and thus, its removal from the dataset is justified."

- Please note that the rubric requires us to mention all the outliers in the dataset. You should be able to identify at least one more outlier in addition to Cheyenne.