

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 is a leading pet store chain in Wyoming with 13 stores throughout the state. with an intent expand and open a 14th store. Data is provided to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Key Decisions:

Answer these questions

1. What decisions needs to be made?

This year, Pawdacity would like to expand and open a 14th store. The requirement is to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

The data provided from which a conclusion needs to be made is as follows:

- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
- A partially parsed data file that can be used for population numbers.
- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

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	343,027.63
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 conducting a Scatterplot on the final dataset, one city outweighs other cities in Total Sales, Gillette = 543,132, and it is a city which skews high in sales, yet does not skew relative to the other data fields. This outlier will be removed 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.