

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?
2. What data is needed to inform those decisions?

ANSWER:

The manager of Pawdacity, a pet store chain is considering of opening its 14th store in Wyoming. We need to recommend where to open the new store based on the data provides. The data provided to us are the historical monthly sales, city/country, population, demographic attributes and so on. This report is to prepare the dataset to perform predictive analysis.

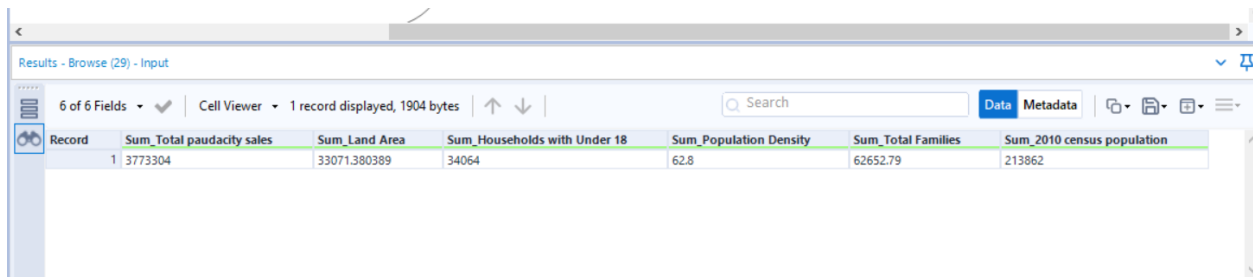
Step 2: Building the Training Set

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

ANSWER

The dataset provided in various CSV files were cleaned. Then we blended and joined the datasets and union them to form a training set. It consist of eleven rows and six columns whose summary is present in the below table.



The screenshot shows a data viewer interface with a table containing 6 columns and 1 row of data. The columns are labeled: Record, Sum_Total paudacity sales, Sum_Land Area, Sum_Households with Under 18, Sum_Population Density, Sum_Total Families, and Sum_2010 census population. The first row contains the values: 1, 3773304, 33071.380389, 34064, 62.8, 62652.79, and 213862.

Record	Sum_Total paudacity sales	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Sum_2010 census population
1	3773304	33071.380389	34064	62.8	62652.79	213862

Column	Sum	Average
Census Population	213,862	213,862
Total Pawdacity Sales	3,773,304	3,773,304
Households with Under 18	34,064	34,064
Land Area	33,071	33,071.38
Population Density	63	62.80
Total Families	62,653	62652.79

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.

ANSWER

I identified the outliers by considering a threshold value which is defined as 1.5 times interquartile range below 1st quartile or above 3rd quartile. It was found that two cities namely Gillette and Cheyenne. Since, Cheyenne had higher number of outliers (i.e. 3 out of 6), I decided to remove Cheyenne.