# 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](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?

*The city for Pawdacity's newest store, based on predicted yearly sales.*

2. What data is needed to inform those decisions?

*Outcome variable: Total Pawdacity Sales (by year).*
*Potential predictors: Census population, household with under 18, land area, population density, total 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* | *19,442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.60* |
| *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.

*Sheridan city is the outlier in the training set. Because the data set is small, I will retain the record of that city and make imputation for all the outliers based on median value. I will not impute the outliers by mean value because standard distribution is not guaranteed.*

## 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.