# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

1. What decisions needs to be made?

    - *I need to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales. But first, I need to identify data needed for it. Thereafter, I have to clean the data which include formatting and blending, and at the end to deal with potential outliers.*

2. What data is needed to inform those decisions?

    - *I had to use partially parsed data file to extract information about 2010 Census population; Monthly sales data for all of the Pawdacity stores for the year 2010; Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families).*

## Step 2: Building the Training Set

| **Column** | **Sum** | **Average** |
|---|---|---|
| *Census Population* | *213,862* | *19,442.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.64* |
| *Households with Under 18* | *34,064* | *3,096.73* |
| *Land Area* | *33,071* | *3,006.45* |
| *Population Density* | *63* | *5.73* |
| *Total Families* | *62,653* | *5,695.73* |

## Step 3: Dealing with Outliers

Are there any cities that are outliers in the training set?

- *I have analyzed all outliers in each field and there are two cities checked as outliers in the training set (Cheyenne and Gillette).*

Which outlier have you chosen to remove or impute?

- *Amongst these two outliers I have decided to keep Gillette because it represents the outlier only in one column (Total Pawdacity Sales), and it’s not necessarily having negative impact on my dataset because of that. The other reason I would rather keep it is that my dataset only consists of 11 cities, and removing all outliers will not be good for further prediction model. On the other hand, I have chosen to remove Cheyenne city from dataset, because it represents the outlier in 4 columns (2010 Census Population, Total Pawdacity Sales, Population Density and Total Families), which drastically skews the line and could lead us to wrong final conclusion and model decision.*