# Data Cleanup

## Project Overview

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th sto
re. The task is to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales. The first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

## Step 1: Business and Data Understanding

1. What decisions needs to be made?
   Depend on the data such as previous total sales and demographic in different cities, choose a city which will generate most profit.

2. What data is needed to inform those decisions?
   City, 2010 census population, total pawdacity sales, household under 18, land area, population density, total families

## Step 2: Building the Training Set

Below results are generated using Ateryx

| **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 |

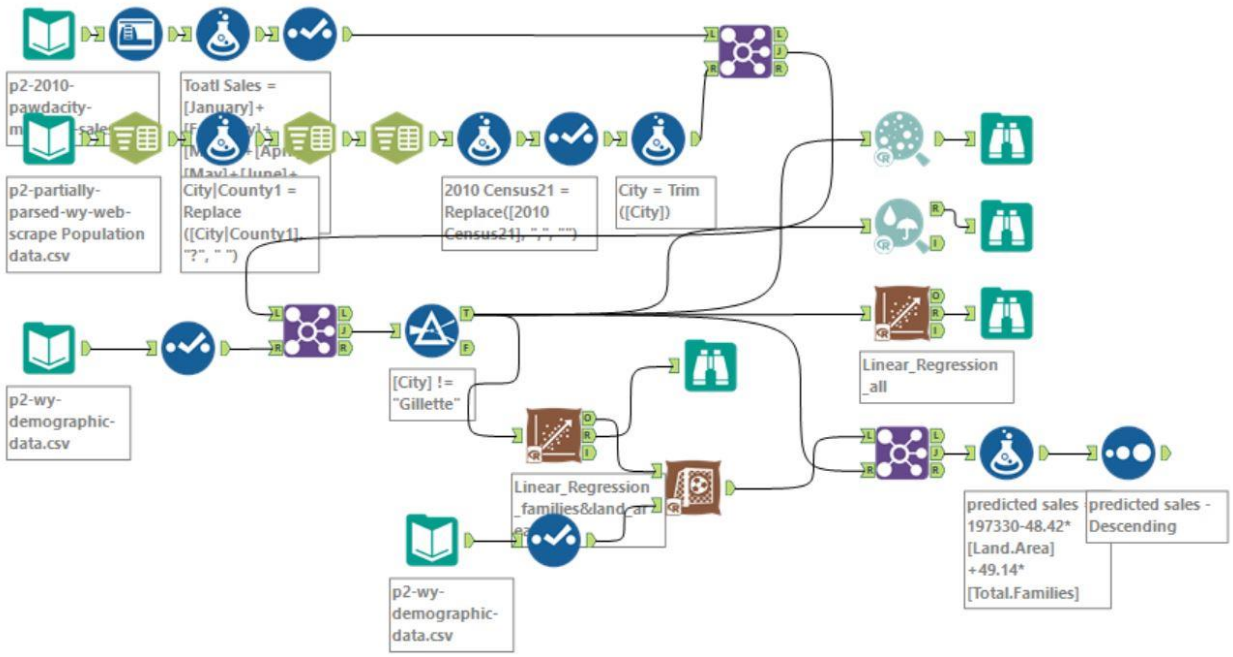


Figure 1. Workflow for building the training set

# Step 3: Dealing with Outliers

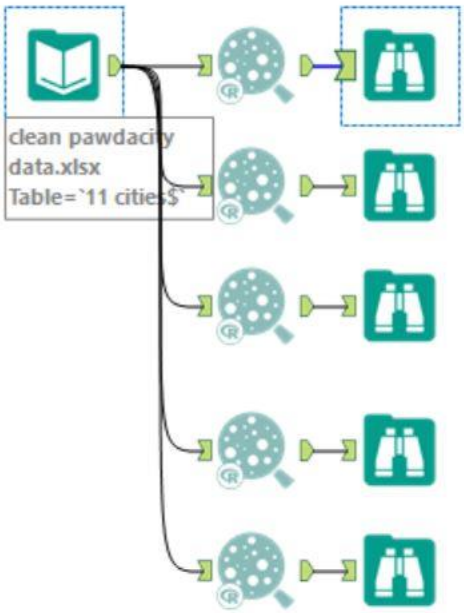


Figure 2. Workflow for identifying outliers

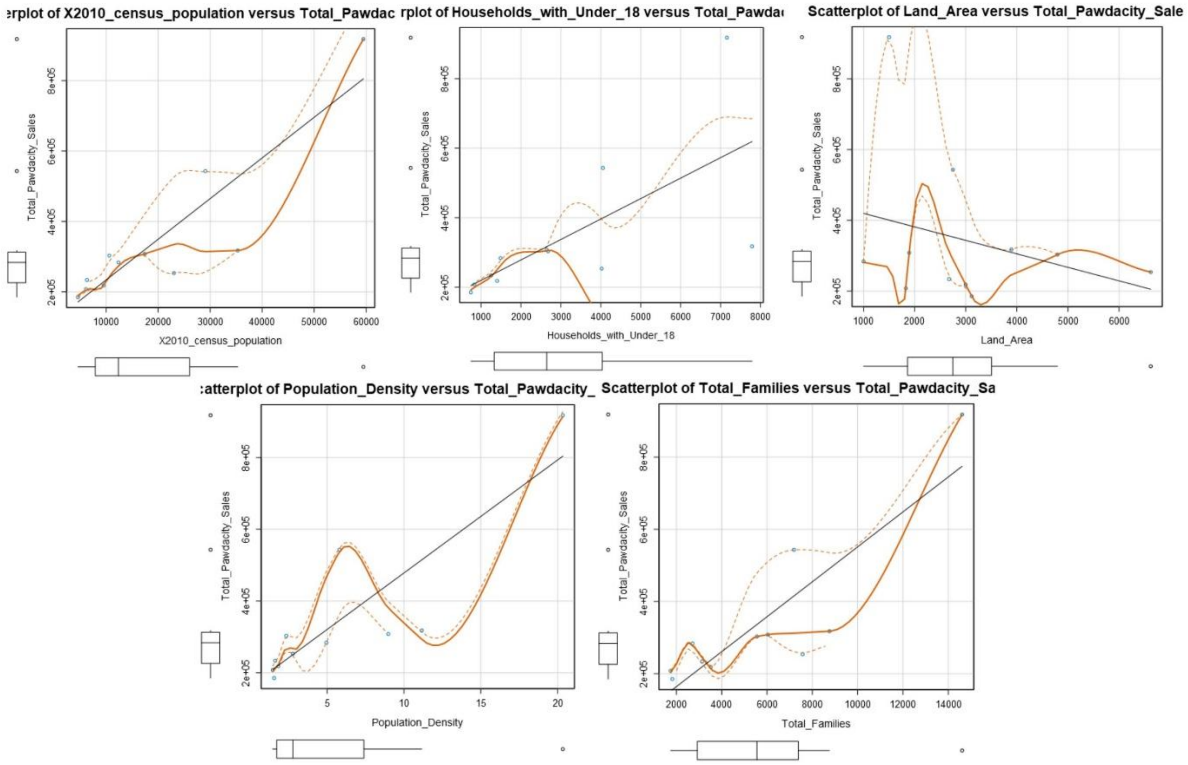


Figure 3. Plots for identifying outliers

After plotting each variable vs total Pawdacity sales for identifying outliers, both city of Cheyenne and Gillette seem like outliers because of high total sales, but when exploring other variables related to Cheyenne's total sales, it look like the high total sales are reasonable. Thus, the outlier is city of Gillette.