# 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?

**This project requires that we clean up and blend the data available from three different data sources. The consolidated data is at the city level and not store level. We would have to analyze and recommend the city for Pawdacity's newest store, based on predicted yearly sales.**

**Consolidated data set consisting of the following:**

**City**
**2010 Census Population**
**Total Pawdacity Sales**
**Households with Under 18**
**Land Area**
**Population Density**
**Total Families**

2. What data is needed to inform those decisions?

**The following data is used to build the consolidated data set:**

- ***p2-2010-pawdacity-monthly-sales.csv*** **- monthly sales for all Pawdacity stores for 2010. (NAME,ADDRESS,CITY,STATE,ZIP)**

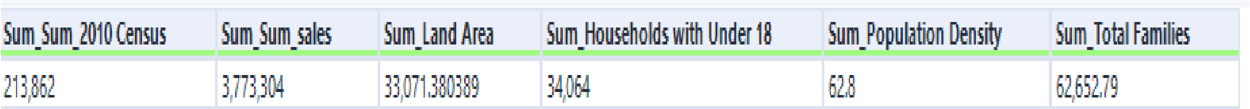
- ***p2-partially-parsed-wy-web-scrape.csv*** **- population numbers. ( City, County, 2014 Estimate, 2010 Census, 2000 Census)**

- *p2-wy-demographic-data.csv* **- demographic data for each city and county in Wyoming. (City, County, Land Area, Households with Under 18, Population Density and Total Families)**

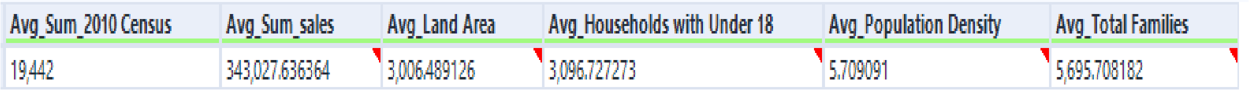
Step 2: Building the Training Set

*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*



| Sum_Sum_2010 Census | Sum_Sum_sales | Sum_Land Area | Sum_Households with Under 18 | Sum_Population Density | Sum_Total Families |
|---|---|---|---|---|---|
| 213,862 | 3,773,304 | 33,071.380389 | 34,064 | 62.8 | 62,652.79 |



| Avg_Sum_2010 Census | Avg_Sum_sales | Avg_Land Area | Avg_Households with Under 18 | Avg_Population Density | Avg_Total Families |
|---|---|---|---|---|---|
| 19,442 | 343,027.636364 | 3,006.489126 | 3,096.727273 | 5.709091 | 5,695.708182 |

| Column | Sum | Average |
|---|---|---|
| *Census Population* | *213,862* | ***19442*** |
| *Total Pawdacity Sales* | *3,773,304* | ***343027.63*** |
| *Households with Under 18* | *34,064* | ***3096.72*** |
| *Land Area* | *33,071* | ***3006.48*** |
| *Population Density* | *63* | ***5.70*** |
| *Total Families* | *62,653* | ***5696*** |

## 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.

**Cheyenne has outliers for 3 out of the 6 variables used in the analysis. Only Cheyenne will be removed because it has outliers in 3 out of 6 data points,also these values adversley effect the analysis. Other cities with outliers are Rock Springs and Gillette and will be retained because they don’t adversley effect the analysis.**