# Creating an Analytical Dataset

## Business and Data Understanding

### Key Decisions:

What decisions needs to be made?

The manager needs to decide which city in Wyoming to expand and open a 14th Pawdacity store based on my analysis for predicted yearly sales.

What data is needed to inform those decisions?

The following files provided by the manager needs to be formatted and blended:

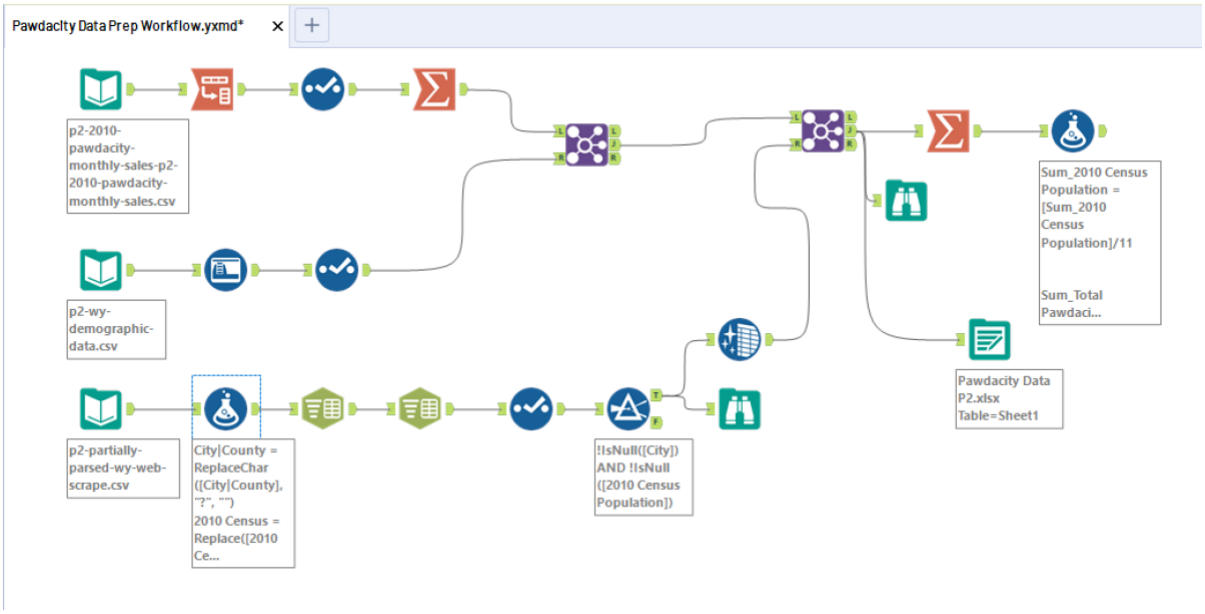
p2-2010-pawdacity-monthly-sales.csv
*p2-wy-demographic-data.csv*
*p2-patially-parsed-wy-web-scrape.csv*

I am specifically interested in analyzing data regarding the City, Census Population, Total Pawdacity Sales, Households with Under 18, Land Area, Population Density, and Total Families in Wyoming.

I chose to omit the NAICS data on most current sales of competitors because we are not particularly concerned with introducing such exogenous variables for the purposes of Pawdacity's future sales forecast model. Such data would be more relevant when considering Pawdacity's susceptibility to market cannibalization.

# Building the Training Set

| Column | Sum | Average |
|---|---|---|
| *Census Population* | *213,862* | *19,422.00* |
| *Total Pawdacity Sales* | *3,773,304* | *343,928.00* |
| *Households with Under 18* | *34,064* | *3,097.00* |
| *Land Area* | *33,071* | *3006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653* | *5695.71* |



# Dealing with Outliers

Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Please explain your reasoning.

There are 2 cities that are outliers in the training set: Cheyenne and Gillette.

**Cheyenne**'s values for **Total Pawdacity Sales**, **Population Density**, and **Total Families** contained points above the upper fence. So, I chose to remove this city from our training set.

**Gillette**'s value for **Total Pawdacity Sales** was slightly above the upper fence for while the values in other columns were within the acceptable range. So, I decided to keep this city in our small dataset.

| City | Total Pawdacity Sales | Land Area | Households with Under 18 | Population Density | Total Families | 2010 Census Population |
|---|---|---|---|---|---|---|
| Buffalo | 185328 | 3115.5075 | 746 | 1.55 | 1819.5 | 4585 |
| Casper | 317736 | 3894.3091 | 7788 | 11.16 | 8756.32 | 35316 |
| Cheyenne | **917892** | 1500.1784 | 7158 | **20.34** | **14612.64** | 59466 |
| Cody | 218376 | 2998.95696 | 1403 | 1.82 | 3515.62 | 9520 |
| Douglas | 208008 | 1829.4651 | 832 | 1.46 | 1744.08 | 6120 |
| Evanston | 283824 | 999.4971 | 1486 | 4.95 | 2712.64 | 12359 |
| Gillette | **543132** | 2748.8529 | 4052 | 5.8 | 7189.43 | 29087 |
| Powell | 233928 | 2673.57455 | 1251 | 1.62 | 3134.18 | 6314 |
| Riverton | 303264 | 4796.859815 | 2680 | 2.34 | 5556.49 | 10615 |
| Rock Springs | 253584 | 6620.201916 | 4022 | 2.78 | 7572.18 | 23036 |
| Sheridan | 308232 | 1893.977048 | 2646 | 8.98 | 6039.71 | 17444 |
| | | | | | | |
| Q1 | 226152 | 1861.721074 | 1327 | 1.72 | 2923.41 | 7917 |
| Q3 | 317736 | 3894.3091 | 4052 | 8.98 | 7572.18 | 29087 |
| IQR | 91584 | 2032.588026 | 2725 | 7.26 | 4648.77 | 21170 |
| Upper Fence | **455112** | 6943.191139 | 8139.5 | **19.87** | **14545.335** | 60842 |
| Lower Fence | 88776 | -1187.16097 | -2760.5 | -9.17 | -4049.745 | -23838 |

Other Considerations:

Due to the ambiguous nature of dealing with outliers, we could conversely make the case that Cheyenne should be included in the training set because it is the capital of Wyoming. Removing the largest and most important city in Wyoming would negatively bias our sales and demographic data so removing Gillette instead would be preserve more data necessary to accurately predict total Pawdacity sales.