# Data Blending and Cleansing

## Business and Data Understanding

**Business Decision**

The business decision that needs to be made is in which city to open another Pawdacity store. Currently Pawdacity has 13 stores and wants to expand to a 14th store. In order to make this decision we need data that is readable and necessary to make the business decision.

**Data Required**

The data required to complete this project is as follows for each city in Wyoming where Pawdacity has at least 1 store location:

Census Population – The size of the city will be important to choosing our next store location as we may get more value out of a larger city

Total Pawdacity Sales – As we are aiming to decide where to open a new location, we will want to use past historical sales in our model to predict future sales.

Households with Under 18 – As we build our model we want multiple predictor variables to determine which are statistically significant. Families with children may be more likely to own pets, and we will use this data to check for significance.

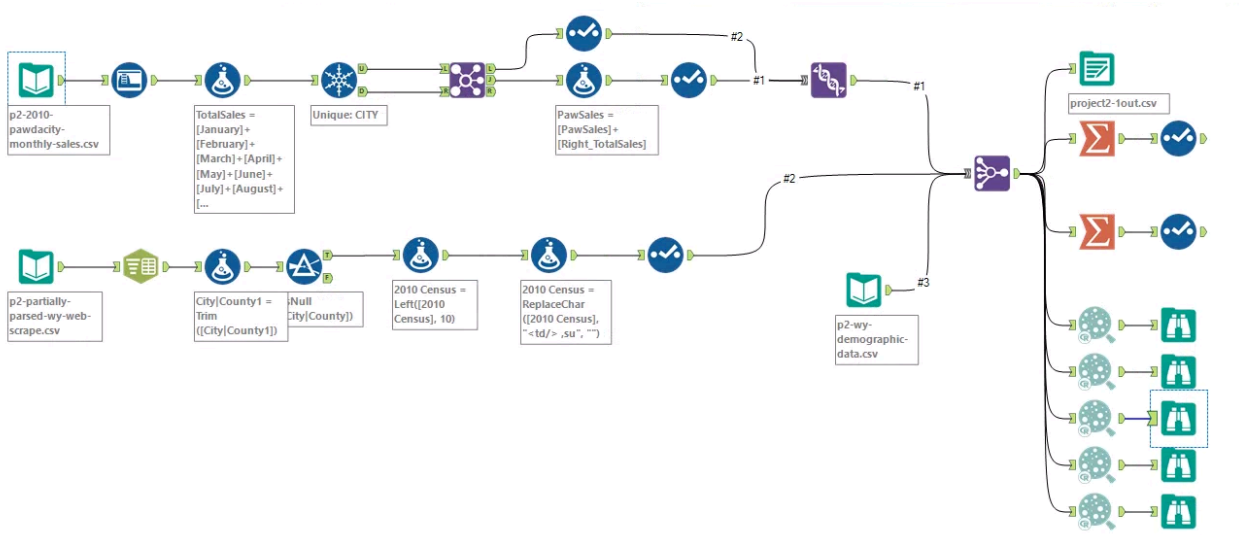
Land Area – We do not want sores that are too close to one another and it could impact the sales if there are stores in close proximity.

Population Density – Population Density may be more significant that population or land area alone and needs to be included.

Total Families – This is another predictor variable that will be needed to check for significance if this has correlation with high sales.

The main goal of this assignment is to blend and clean the data given, so that we can extract only the pertinent data, and so that it is readable and usable in future calculations and decisions.

## Building the Training Set




| Record # | Sum_2010Census | Sum_PawSales | Sum_Households with Under 18 | Sum_Land Area | Sum_Population Density | Sum_Total Families |
|---|---|---|---|---|---|---|
| 1 | 213862 | 3773304 | 34064 | 33071 | 63 | 62653 |

| Record # | Avg_2010Census | Avg_PawSales | Avg_Households with Under 18 | Avg_Land Area | Avg_Population Density | Avg_Total Families |
|---|---|---|---|---|---|---|
| 1 | 19442.00 | 343027.64 | 3096.73 | 3006.49 | 5.71 | 5695.71 |

| Column | Sum | Average |
|---|---|---|
| Census Population | 213,862 | 19442.00 |
| Total Pawdacity Sales | 3,773,304 | 343027.64 |
| Households with Under 18 | 34,064 | 3096.73 |
| Land Area | 33,071 | 3006.49 |
| Population Density | 63 | 5.71 |
| Total Families | 62,653 | 5695.71 |

Four data files were included for consideration, or which, only three were needed for output.

p2-2010-pawdacity-monthly-sales.csv – includes monthly sales of each Pawdacity store. This data needed to be converted to yearly sales, and on a city by city basis rather than store by store. First, a new column, TotalSales, was created to sum the sales from each month. Then, two cities contained two stores and needed to be merged through a “Unique” tool and then sales merged together.

p2-partially-parsed-wy-web-scrape.csv – Includes population data, which was scraped from the internet. This data needed to be condensed and cleaned. First the City|County column needed to be split with the “Text to Columns” tool then cleaned with a Trim formula and a isNull function. Also, the extra characters needed to be removed from the 2010 Census data with the Left and ReplaceChar formulas.

p2-wy-demographic-data.csv – Contains family size, both 'households with under 18' and 'total families,' as well as 'population density.' This data was rather clean, the field type needed to be changed from a string to a double in order to use in calculations.

After the data was cleaned, they were blended to create an output of only necessary information. After which, the figures were summed and scatterplots created to check for accuracy and decision making.

## Dealing with Outliers

| CITY | 2010Census | PawSales | Households with | Land Area | Population Density | Total Families |
|---|---|---|---|---|---|---|
| Douglas | 6120 | 208008 | 832 | 1829.4651 | 1.46 | 1744.08 |
| Buffalo | 4585 | 185328 | 746 | 3115.5075 | 1.55 | 1819.5 |
| Evanston | 12359 | 283824 | 1486 | 999.4971 | 4.95 | 2712.64 |
| Powell | 6314 | 233928 | 1251 | 2673.57455 | 1.62 | 3134.18 |
| Cody | 9520 | 218376 | 1403 | 2998.95696 | 1.82 | 3515.62 |
| Riverton | 10615 | 303264 | 2680 | 4796.859815 | 2.34 | 5556.49 |
| Sheridan | 17444 | 308232 | 2646 | 1893.977048 | 8.98 | 6039.71 |
| Gillette | 29087 | 543132 | 4052 | 2748.8529 | 5.8 | 7189.43 |
| Rock Springs | 23036 | 253584 | 4022 | 6620.201916 | 2.78 | 7572.18 |
| Casper | 35316 | 317736 | 7788 | 3894.3091 | 11.16 | 8756.32 |
| Cheyenne | 59466 | 917892 | 7158 | 1500.1784 | 20.34 | 14612.64 |
| | | | | | | |
| | | | | | | |
| Median | 12359 | 283824 | 2646 | 2748.8529 | 2.78 | 5556.49 |
| 1st Quartile | 7917 | 226152 | 1327 | 1861.721074 | 1.72 | 2923.41 |
| 3rd Quartile | 26061.5 | 312984 | 4037 | 3504.9083 | 7.39 | 7380.805 |
| IQR | 18144.5 | 86832 | 2710 | 1643.187226 | 5.67 | 4457.395 |
| IQR * 1.5 | 27216.75 | 130248 | 4065 | 2464.780839 | 8.505 | 6686.0925 |
| Upper Fence | 53278.25 | 443232 | 8102 | 5969.689139 | 15.895 | 14066.8975 |
| Lower Fence | -19299.75 | 95904 | -2738 | -603.059765 | -6.785 | -3762.6825 |

After determining the interquartile range, we can see which values are outliers.

Cheyenne – has four values that are outliers: 2010 census, Pawdacity sales, population density, and total families. More money has been spent in Cheyenne even though it has low land area and households with children.

Gillette – has one value that is an outlier, Pawdacity Sales. This and Cheyenne both have two stores.

Rock Springs – has one value that is an outlier, land area.

Because most cities are not the size of Cheyenne, and with such high density, this is the city that I would choose to exclude from the model. Most of the cities we will run this model against, Cheyenne would skew those results, as such, we will have more accurate results through excluding Cheyenne. The other two cities with outliers are Gillette and Rock Springs, both of these cities are a close to the same density of other cities in the model, they are within one standard deviation. By removing too many cities from the model, the model becomes less reliable, and we should be hesitant to remove many cities. It is possible there are other cities that the model is tested against that will compare to both Gillette and Rock Springs, therefore by including these cities, our model is more accurate. Both Gillette and Rock Springs will remain in the model, and Cheyenne will be removed from the model.