# Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

1. What decisions needs to be made?

I need to recommend the city for Pawdacity's newest store based on the predicted yearly sales to find the best new location.

2. What data is needed to inform those decisions?

- Demographic data such as City, County land area, household, population, total families of current stores
- Population data
- Previous monthly sales data to base our prediction on yearly sale data

## Step 2: Building the Training Set

The following is the sum and average of the cleaned and newly blended data set calculated through using the Alteryx platform

| Column | Sum | Average |
|---|---|---|
| *Census Population* | *213,862* | 19,442 |
| *Total Pawdacity Sales* | *3,773,304* | 343027.636363636 |
| *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 |

Table 1: Data Set Sum & Average Calculation

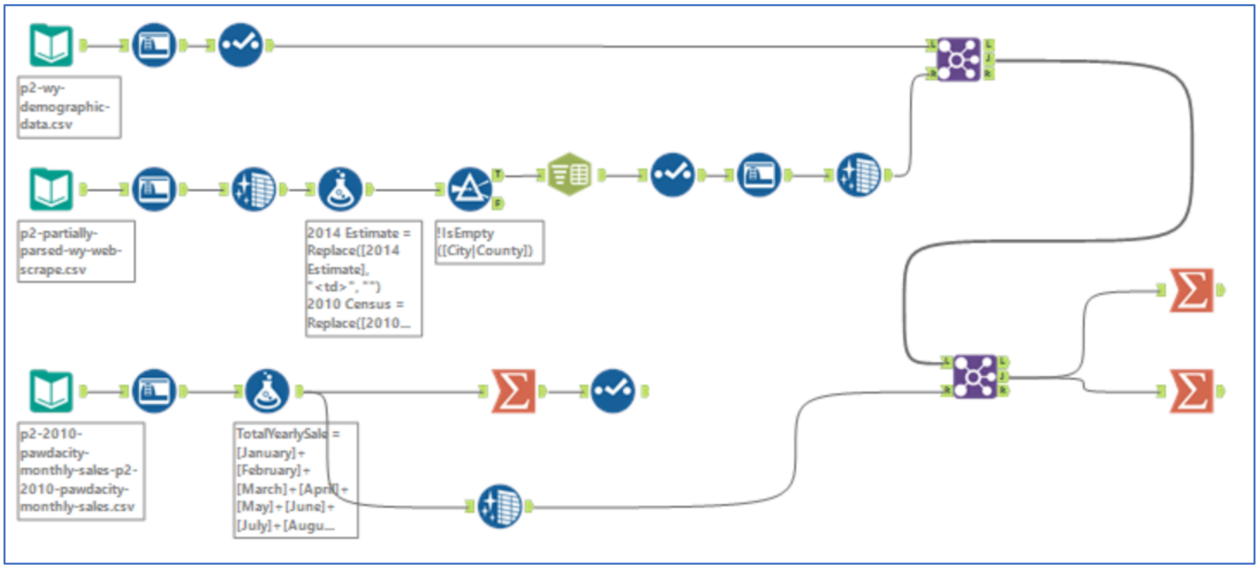
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
2014 Estimate = Replace([2014 Estimate], "<td>", "") 2010 Census = Replace([2010...
!IsEmpty ([City|County])
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
TotalYearlySale = [January]+ [February]+ [March]+[April]+ [May]+[June]+ [July]+[Augu...

Figure 1: Alteryx data cleaning, formatting, and blending workflow

## 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?

After I created the clean data set, to figure out the outlier, I have used scatterplots and calculated the IQR to investigate.

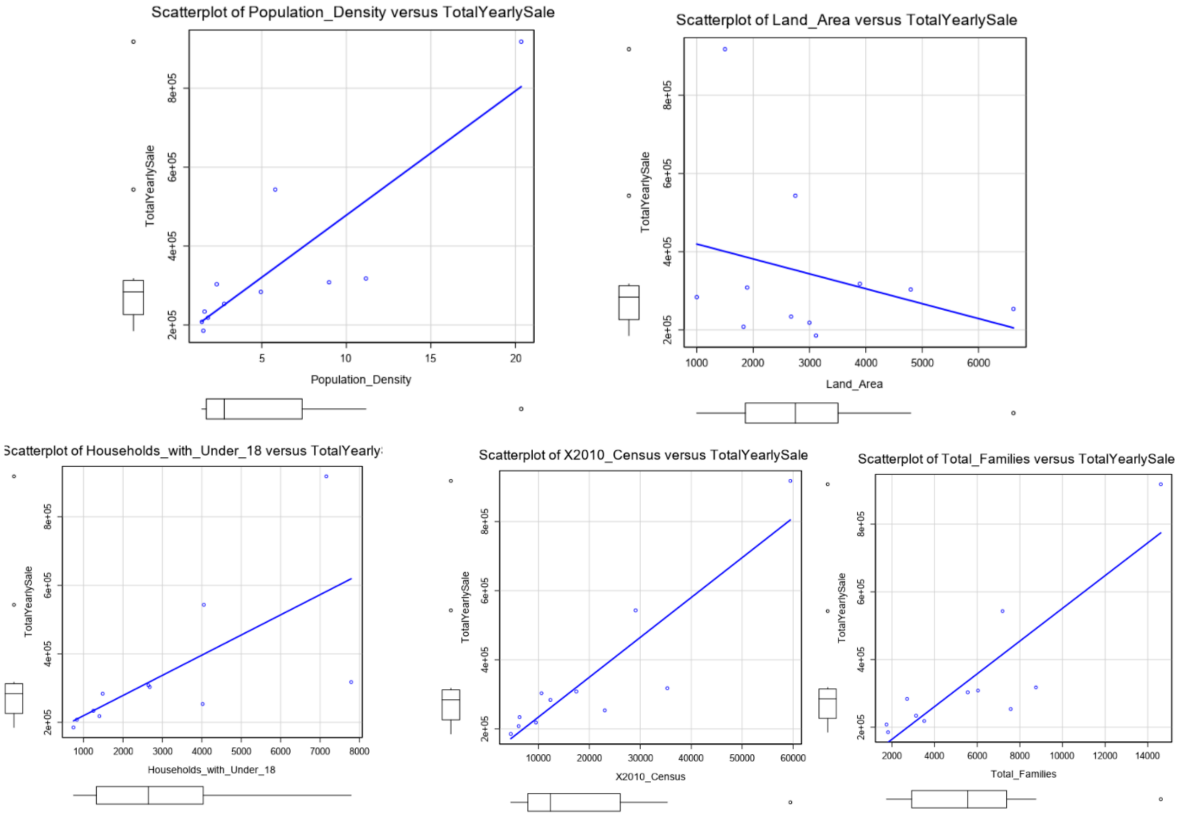

Figure 1: Scatterplots to check for outliers

Calculating IQR to find outliers

| City | Land Area | Households with Under 18 | Population Density | Total Families | TotalYearlySale | 2010 Census |
|---|---|---|---|---|---|---|
| Buffalo | 3115.5075 | 746 | 1.55 | 1819.5 | 185,328 | 4585 |
| Casper | 3894.3091 | 7788 | 11.16 | 8756.32 | 317,736 | 35316 |
| Cheyenne | 1500.1784 | 7158 | 20.34 | 14612.64 | 917,892 | 59466 |
| Cody | 2998.95696 | 1403 | 1.82 | 3515.62 | 218,376 | 9520 |
| Douglas | 1829.4651 | 832 | 1.46 | 1744.08 | 208,008 | 6120 |
| Evanston | 999.4971 | 1486 | 4.95 | 2712.64 | 283,824 | 12359 |
| Gillette | 2748.8529 | 4052 | 5.8 | 7189.43 | 543,132 | 29087 |
| Powell | 2673.57455 | 1251 | 1.62 | 3134.18 | 233,928 | 6314 |
| Riverton | 4796.859815 | 2680 | 2.34 | 5556.49 | 303,264 | 10615 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Rock Springs | 6620.201916 | 4022 | 2.78 | 7572.18 | 253,584 | 23036 |
| Sheridan | 1893.977048 | 2646 | 8.98 | 6039.71 | 308,232 | 17444 |
| Median | 2748.8529 | 2646 | 2.78 | 5556.49 | 283,824 | 12359 |
| 1st quartile | 1829.4651 | 1251 | 1.62 | 2712.64 | 218376 | 6314 |
| 3rd quartile | 3894.3091 | 4052 | 8.98 | 7572.18 | 317736 | 29087 |
| Interquartile Range: IQR = Q3 - Q1 | 2064.844 | 2801 | 7.36 | 4859.54 | 99360 | 22773 |
| Upper Fence = Q3 + 1.5 * IQR | 6991.5751 | 8253.5 | 20.02 | 14861.49 | 466776 | 63246.5 |
| Lower Fence = Q1 - 1.5 * IQR | -1267.8009 | -2950.5 | -9.42 | -4576.67 | 69336 | -27845.5 |

Table 2: IQR calculation from Excel

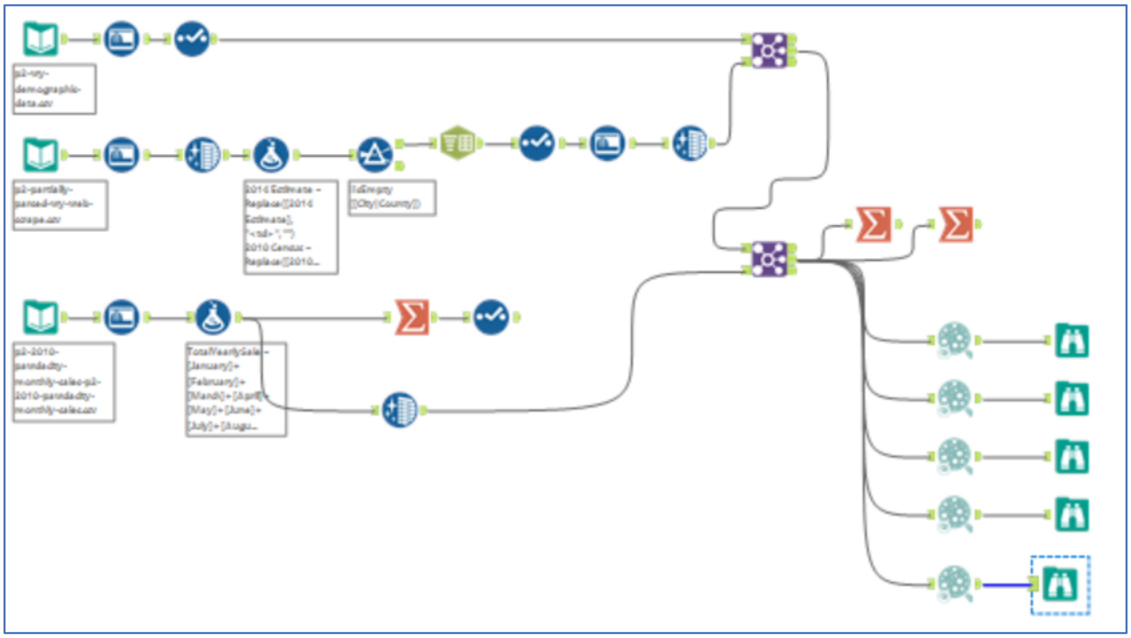
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
2014 Estimate = Replace([2014 Estimate], "<td>", "") 2010 Census = Replace([2010...
!IsEmpty([City|County])
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
TotalYearlySales = [January]+ [February]+ [March]+[April]+ [May]+[June]+ [July]+[Augu...

Figure 3: Alteryx workflow after adding scatterplot and data brows to check for outliers

Based on the scatter plots and IQR calculations, Cheyenne and Gillette showed signs of being outliers (highlighted in yellow in Table 2) when it came to the total yearly sales. I will amputate **Cheyenne** as it has showed significant amount of data in terms of having almost double the yearly sales, population density, and total families. I will remove it because it could give us false data if kept. Also, I will keep all the other cities as the data set is considered short and only consists of 11 rows.