Data Cleanup

Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

1. What decisions needs to be made?

Pawdacity, a leading pet store chain in Wyoming, need to decide where to open its 14th store.

2. What data is needed to inform those decisions?

Data required in order to inform this decision are city, 2010 census population, Pawdacity sales in other stores, competitor sales, household with under 18, land area, population density and total families.

Building the Training Set

By performing the select, formula, data cleansing and filter functions on 4 datasets, the averages for the variables below were obtained. Also attached below is the workflow to obtain the averages.

Table 1: Sums and Averages of Variables

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

Dealing with Outliers

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.

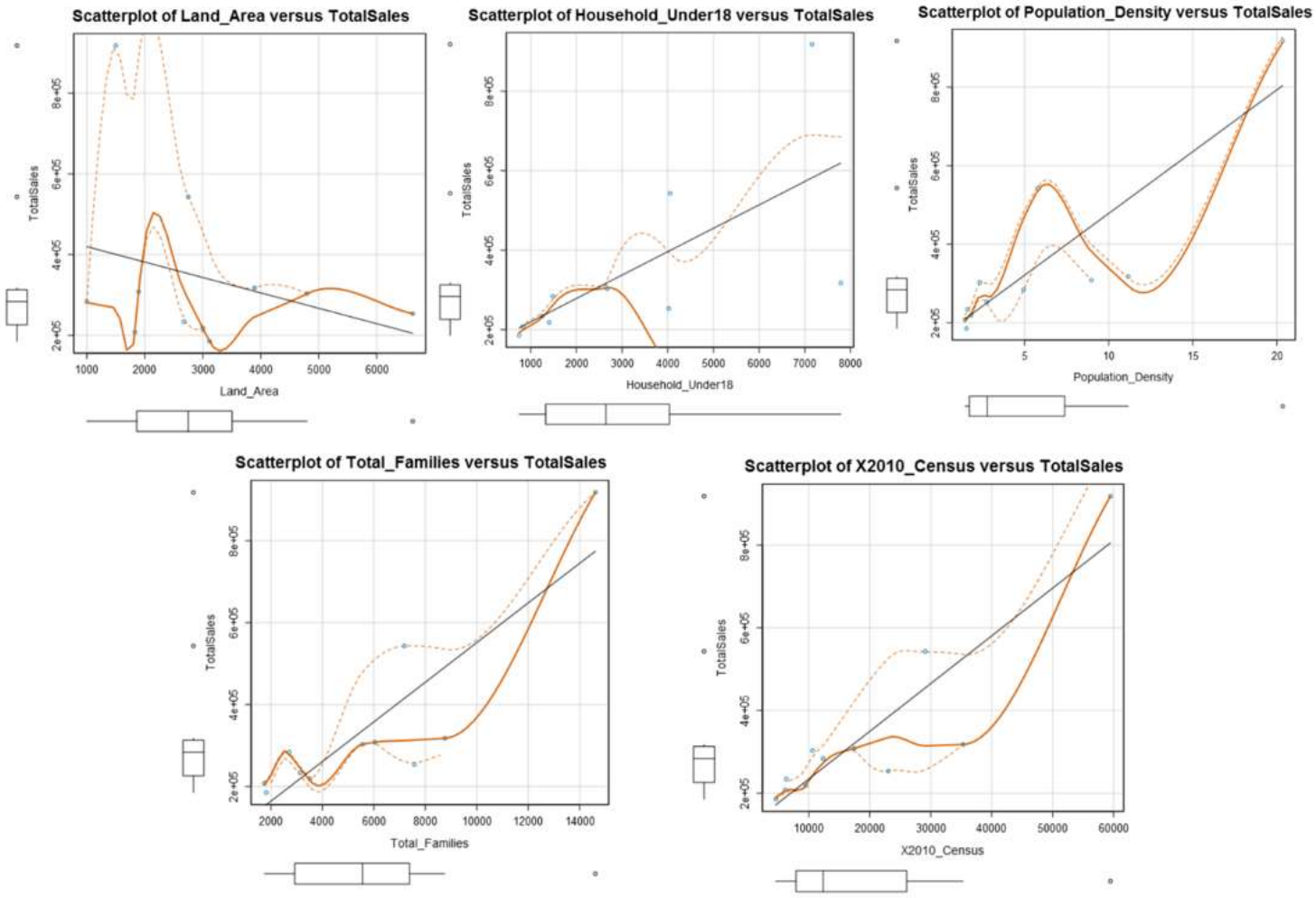

Figure 1: Scatterplots of Population-related variables versus Pawdacity Total Sales

Based on the 5 scatterplots above, the city of **Gillette** and **Cheyenne*** seem to be the outliers as their sales data are higher than expected.
When the scatterplots are extrapolated, Cheyenne's sales data falls within the expected range when extrapolated.
Thus, Gillette would be the outlier in this case when compared against all other cities due to its greatest distance from the linear trend.
Since the relationships between Gillette's population related variables and total sales are still correlated, Gillette should be kept for analysis.

Alteryx Workflow

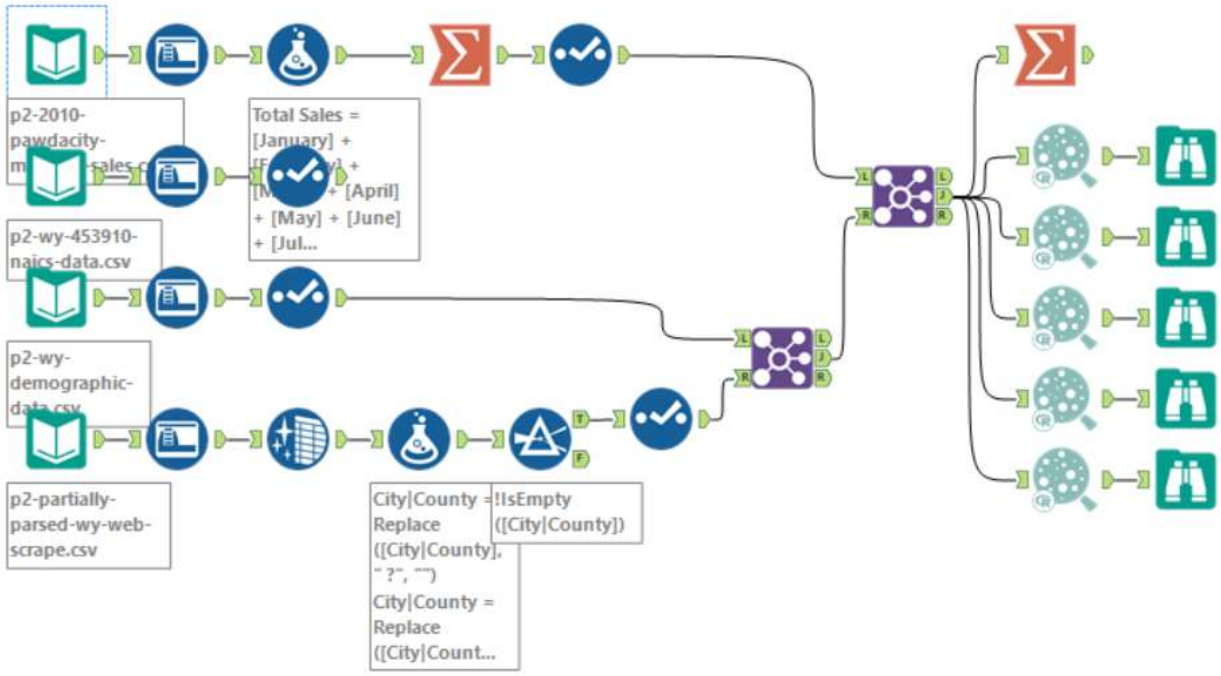

Figure 2: Workflow to obtain sums and averages of variables