Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

### Key Decisions:

1. **What decisions needs to be made?**

Which city is the best for opening a 14th store?

1. **What data is needed to inform those decisions?**

- The monthly sales data for all of the Pawdacity stores for the year 2010.

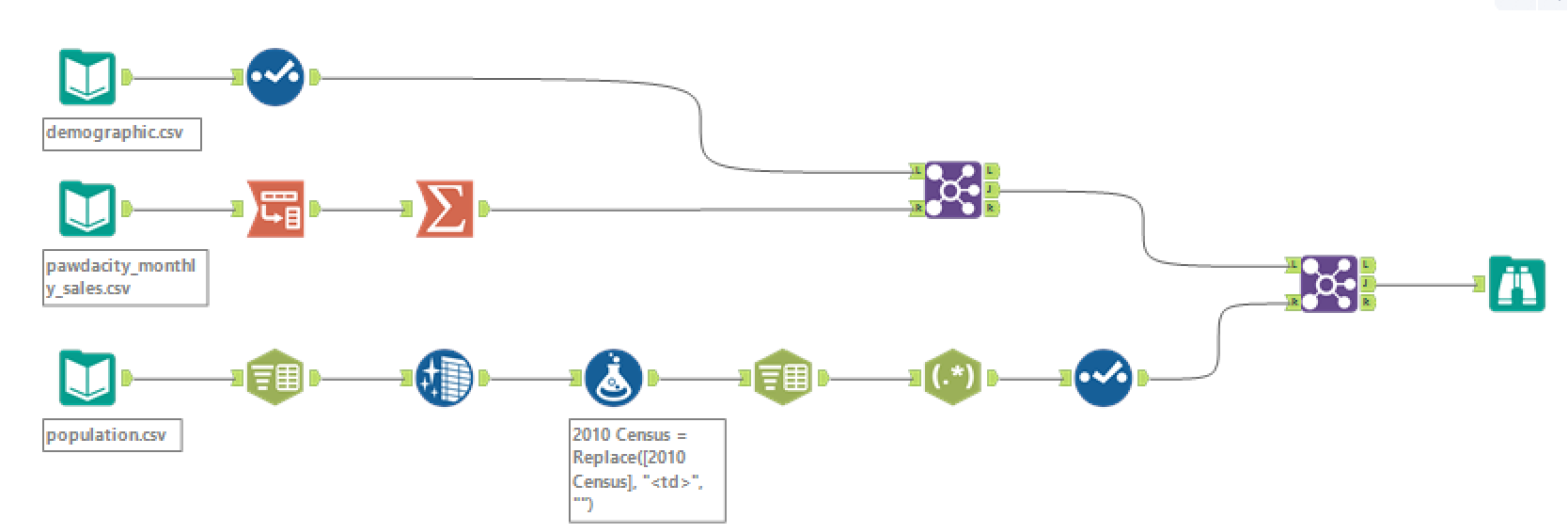
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.

- A partially parsed data file that can be used for population numbers.

- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyomin.

## Step 2: Building the Training Set

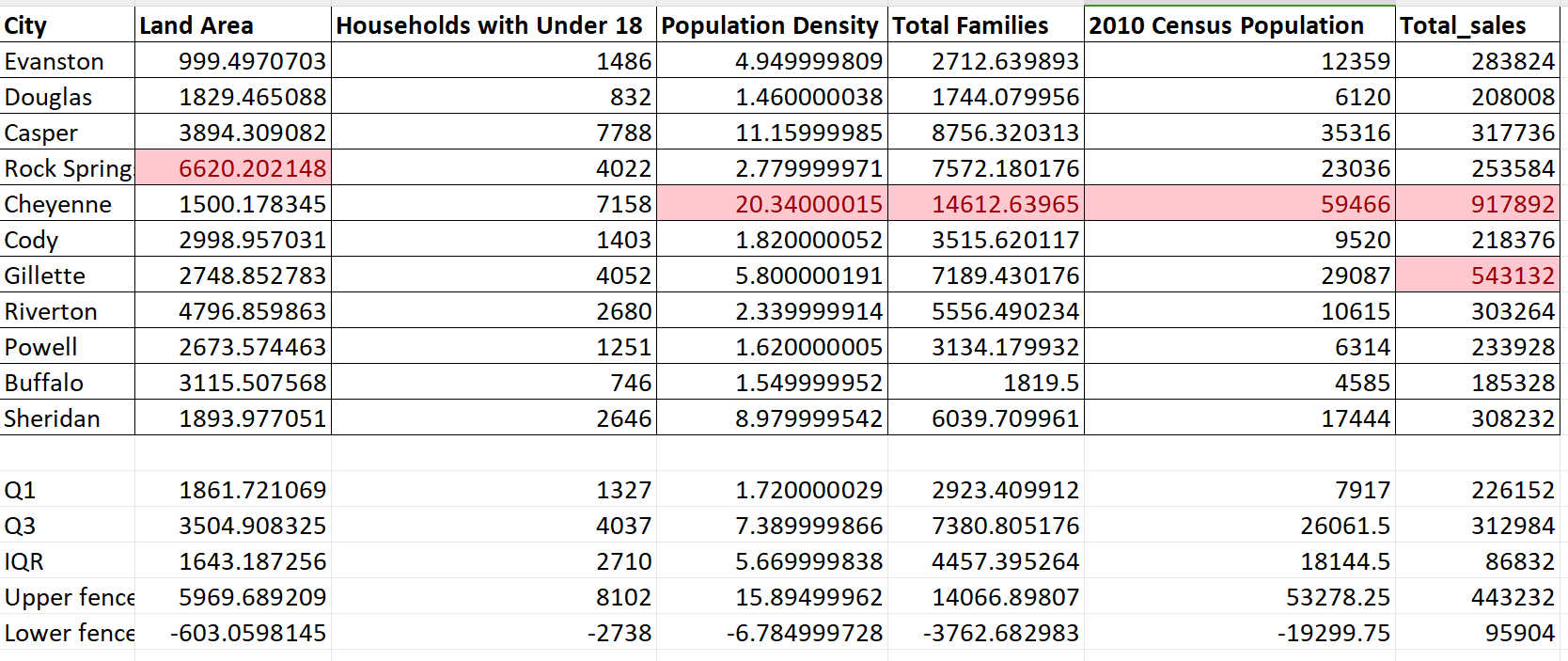
This is the workflow performed in Alteryx.



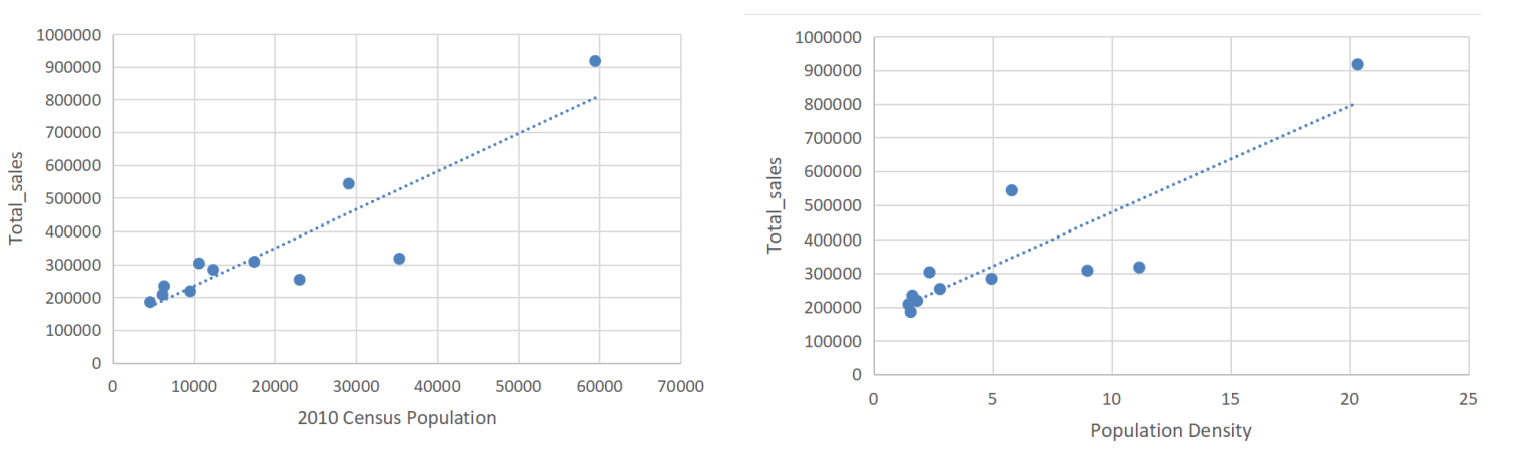
|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19,442* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.64* |
| *Households with Under 18* | *34,064* | *3,096.73* |
| *Land Area* | *33,071* | *3,006.49* |
| *Population Density* | *63* | *5.71* |
| *Total Families* | *62,653* | *5,695.71* |

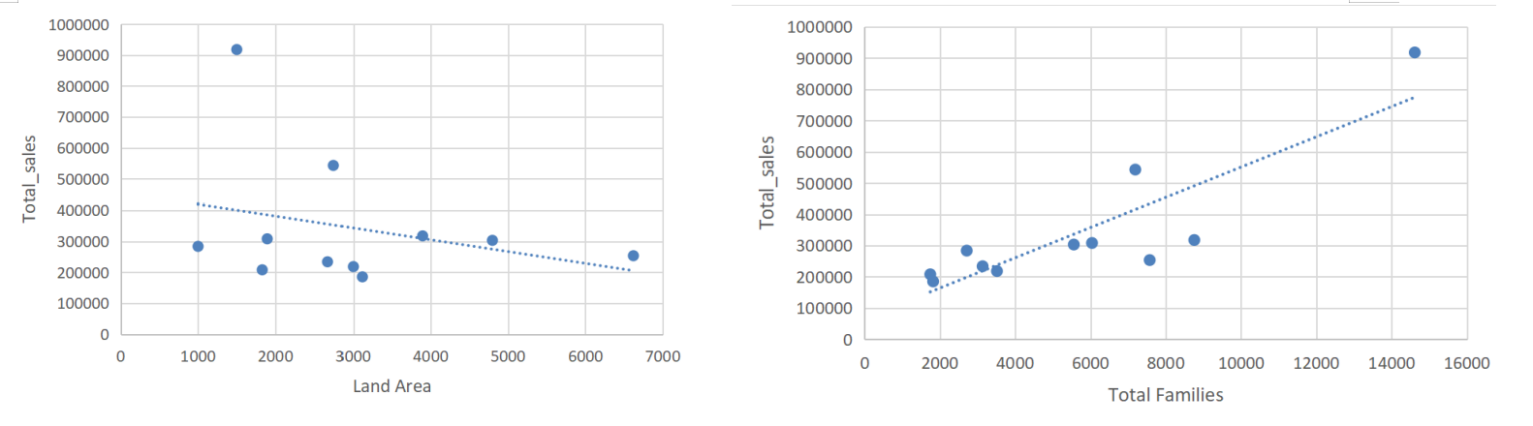
## Step 3: Dealing with Outliers

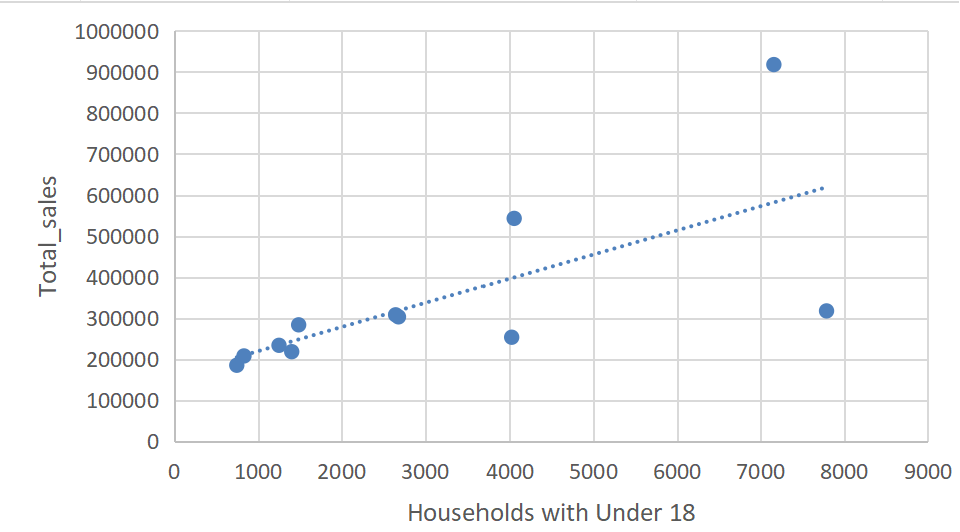
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.



We can see that values marked red with pink background are outliers in each variables, then find matching cities (Rock Springs, Cheyenne, and Gillette) that were outliers in the training set. Using scatter plot to investigate the relationship between each predictor with the variable “Total\_sales” as below:







We can see that Cheyenne has outliers in most of variables. Doing a small research and I find that Cheyenne is the capital and most populous city of the U.S. state of Wyoming. It makes sense for Cheyenne to be different from other cities in Wyoming. Although Cheyenne has second smallest land area, this city has the largest total sales.

Therefore, I will remove the city Cheyenne from the training set to avoid biased model due to the effect of outliers. See the scatter plot showing the relationship between land area and total sales below, now the slope of the line is increased after removing outlier.

