Project 2.1: Data Cleanup

## Step 1: Business and Data Understanding

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

### Key Decisions:

*Answer these questions*

1. What decisions needs to be made?

Pawdacity currently has 13 stores in Wyoming. The decision that needs to be made is; which city should the new store be located in.

1. What data is needed to inform those decisions?

The data needed to inform this decision is as followed at a city level:

* Total sales of current Pawdacity stores.
* Current population estimations for cities in Wyoming.
* Current demographic data for populations in Wyoming at a city level including; Total Families, Households with under 18, Population density.
* Land area of cities in Wyoming.
* Total sales of competitors.

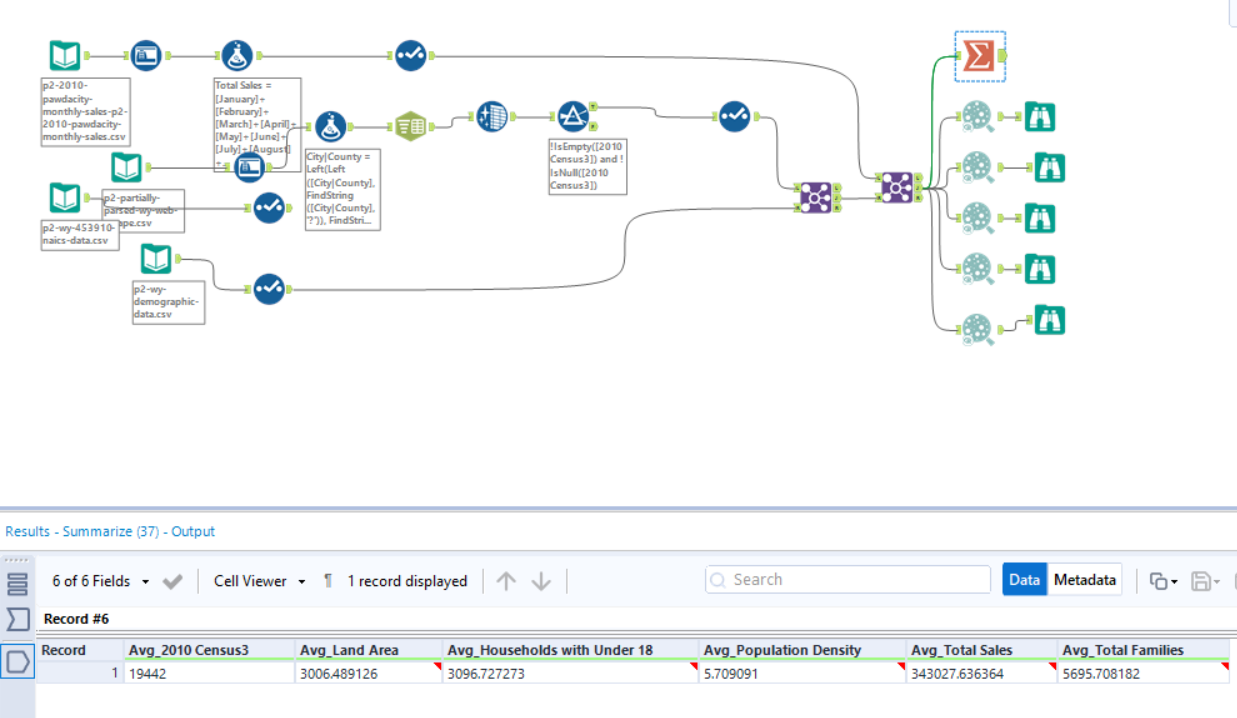
## Step 2: Building the Training Set

*Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.*

*In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24*

|  |  |  |
| --- | --- | --- |
| **Column** | **Sum** | **Average** |
| *Census Population* | *213,862* | *19442* |
| *Total Pawdacity Sales* | *3,773,304* | *343,027.64* |
| *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* |

*Alteryx Workflow for cleaning the data below:*

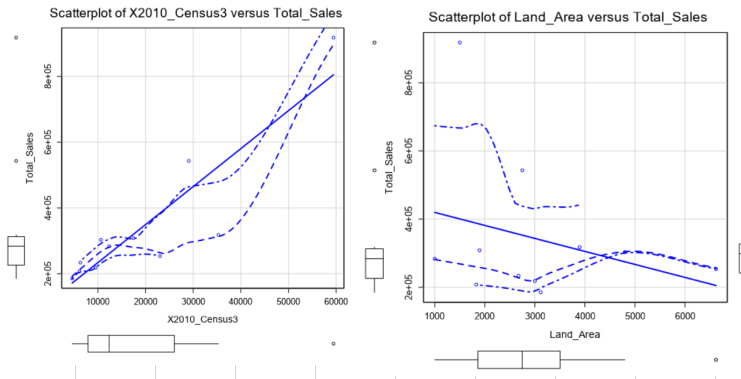


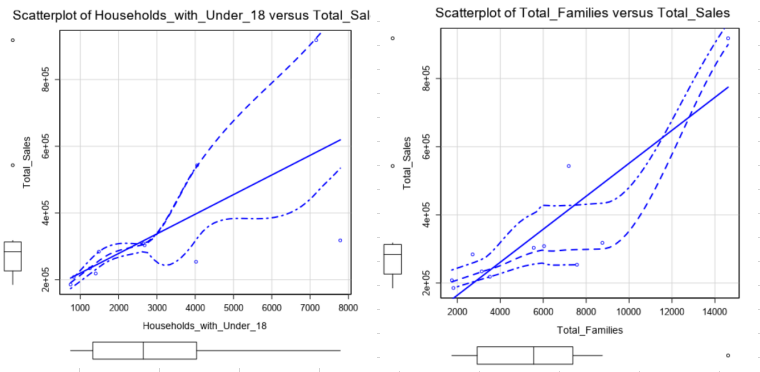
## Step 3: Dealing with Outliers

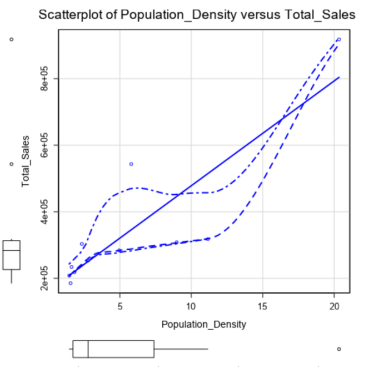
*Answer these questions*

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

I will remove the city of **Cheyenne** as an outlier based on the following analysis.







The following is an analysis of the outliers in each of the numeric variables:



IQR Outlier Table

(False means outside upper and lower fence base on IQR Method)



As seen above, Cheyenne is an outlier in four out of the six numeric variables, where as only two other cities are an outlier in one numeric variable each. Although Cheyenne does seem to follow the trend if we extrapolate the linear trend-line in a couple of them, it is still the most appropriate to remove.