Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. **What decisions needs to be made?**
   - i) The company wants to analyse the data for prediction of new pet store in the Pawdacity.
   - ii) Check if there is any missing or duplicate data in any of the columns.
   - iii) Fields and tools to be used needs to be decided.

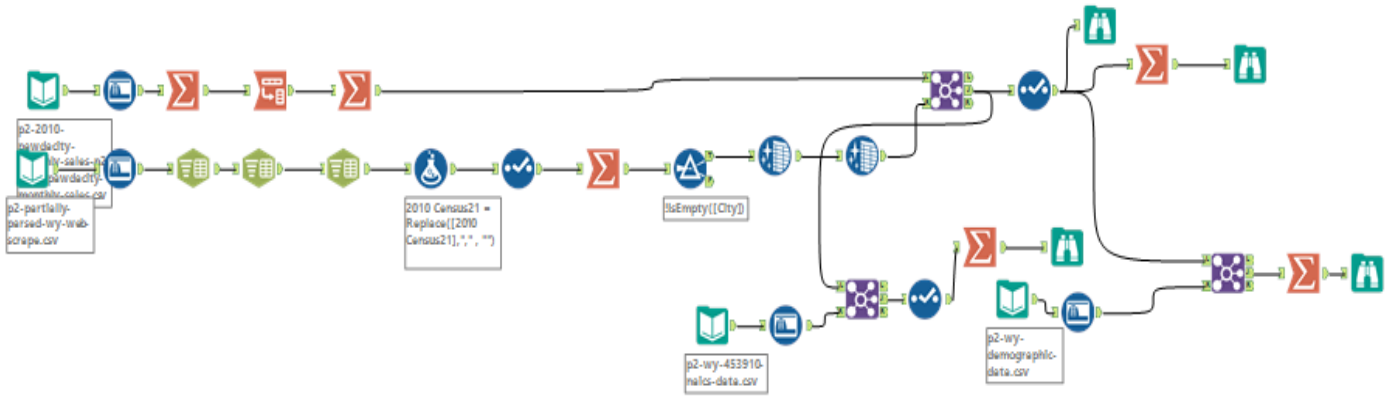
2. **What data is needed to inform those decisions?**
   The data required in order to inform these decision are *city*, *2010 census population*, *Pawdacity sales in other stores*, *competitor sales*, *household with under 18*, *land area*, *population density* and *total families*.

Step 2: Building the Training Set

By performing the select, formula, data cleansing and filter functions on datasets, the averages for the variables below were obtained.

Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.63*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Alteryx Workflow:



Step 3: Dealing with Outliers

We have two cities which have outliers.

1. Cheyenne
2. Gillette

That outlier will be removed which has more potential and causes considerable amount of change to the results and the predictive model.

So, according to the above mentioned reason, we will remove **Cheyenne** from the data set.