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, a leading pet store chain in Wyoming, needs recommendation on where to open its 14th store. Some of the data required in order to inform this decision are *city*, *2010 census population*,

2. What data is needed to inform those decisions?
   Pawdacity sales in other stores, *competitor sales*, *household with under 18*, *land area*, *population density* and *total families*.

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*	343027.64
Households with Under 18	*34,064*	*18335*
Land Area	*33,071*	*42074*
Population Density	*63*	*22*
Total Families	*62,653*	*40.5*

Step 3: Dealing with Outliers

Answer these questions

I use excel sheet to deal with the outliers , there are several outliers I chose Households with Under 18

To calculate the upper fence and the lower fence, here are the exact steps:

1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. Excel function QUARTILE.INC or QUARTILE.EXC

Q1=54.5

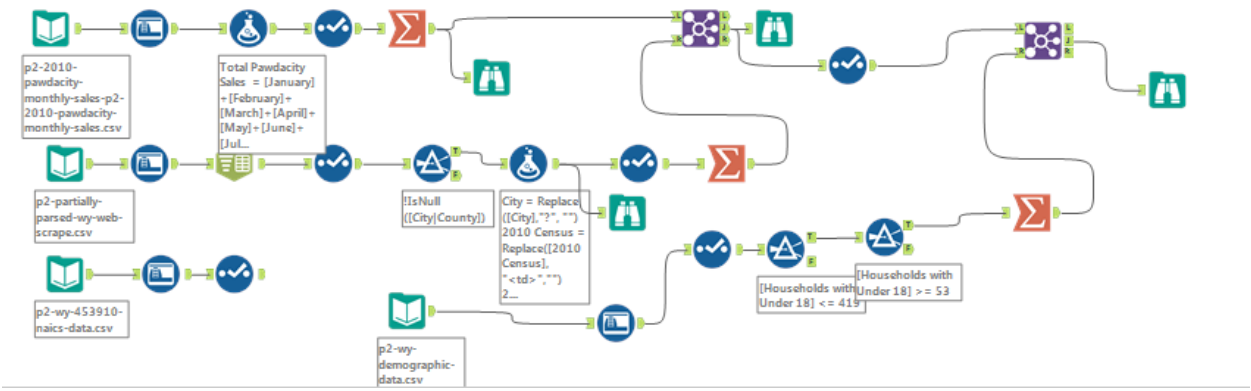
Q3=417

2 . Calculate the Interquartile Range: IQR = Q3 - Q1=417-54.5=362.5

3 . Add 1.5 *IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5* IQR
=418.5

4 . Subtract 1.5 *IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5* IQR
=53

Alteryx Workflow



.Figure 1: Workflow to obtain sums and averages of variables