Project 2.1: Data Cleanup

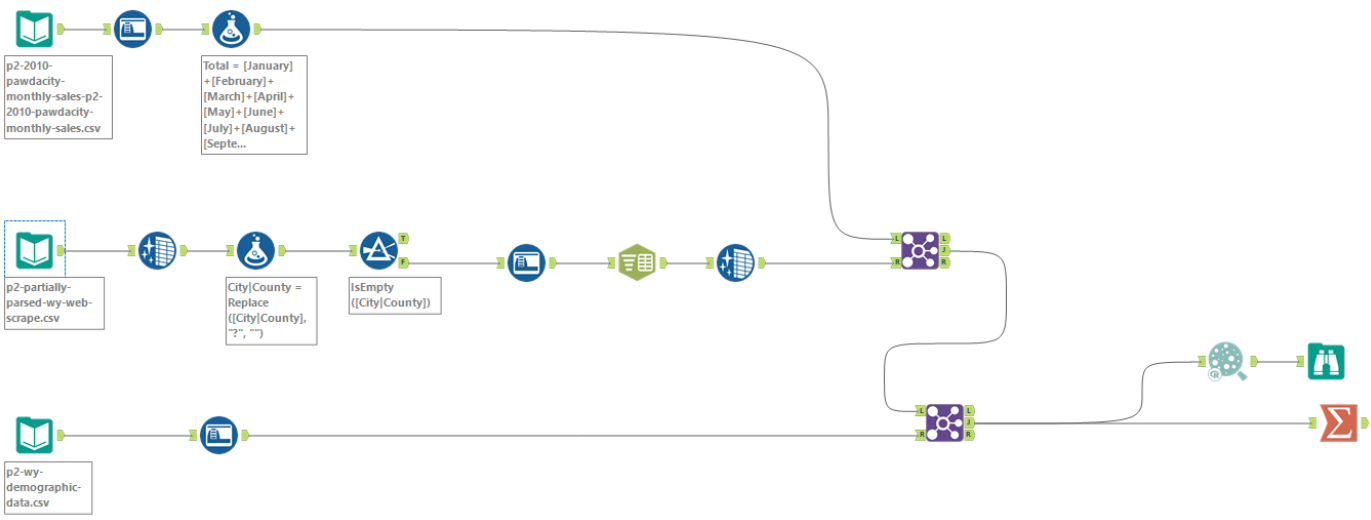
Step 1: Business and Data Understanding

1. What decisions needs to be made?
   Pawdacity, which is a leading pet store chain in Wyoming with 13 stores, would like to expand and open a 14th store. We need to perform analysis to recommend the city for Pawdacity's newest store based on predicted yearly sales.

2. What data is needed to inform those decisions?
   Based on existing data:
   -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 Wyoming.
   We will use the data to extract a data set for the following columns :(City,2010 Census Population,Total Pawdacity Sales,Households with,Under 18,Land Area,Population Density,Total Families) to make the right decision about the new store.

Step 2: Building the Training Set

Workflow alteryx:



Output after cleaning the data:

Record	CITY	Total	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
2	Casper	317736	35316	3894.3091	7788	11.16	8756.32
3	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
4	Cody	218376	9520	2998.95696	1403	1.82	3515.62
5	Douglas	208008	6120	1829.4651	832	1.46	1744.08
6	Evanston	283824	12359	999.4971	1486	4.95	2712.64
7	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
8	Powell	233928	6314	2673.57455	1251	1.62	3134.18
9	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
10	Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
11	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

* Total = Total Pawdacity Sales

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)?

IQR Steps:

1 . Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset.
2 . Calculate the Interquartile Range: IQR = Q3 - Q1
3 . Add 1.5 *IQR to Q3 to get the upper fence: Upper Fence = Q3 + 1.5* IQR
4 . Subtract 1.5 *IQR to Q1 to get the lower fence: Lower Fence = Q1 - 1.5* IQR
5 . Values above the Upper Fence and values below the Lower Fence are outliers

The result:

City	2010 Census Population	Total Pawdacity Sales	Households with Under 18	Land Area	Popuation Density	Total Families
Buffalo	4585	185328	746	3115.508	1.55	1819.5
Casper	35316	317736	7788	3894.309	11.16	8756.32
Cheyenne	59466	917892	7158	1500.178	20.34	14612.64
Cody	9520	218376	1403	2998.957	1.82	3515.62
Douglas	6120	208008	832	1829.465	1.46	1744.08
Evanston	12359	283824	1486	999.4971	4.95	2712.64
Gillette	29087	543132	4052	2748.853	5.8	7189.43
Powell	6314	233928	1251	2673.575	1.62	3134.18
Riverton	10615	303264	2680	4796.86	2.34	5556.49
Rock Springs	23036	253584	4022	6620.202	2.78	7572.18
Sheridan	17444	308232	2646	1893.977	8.98	6039.71
Outliers						
Median	12359	283824	2646	2748.853	2.78	5556.49
1st Quartile	6314	218376	1251	1829.465	1.62	2712.64
3rd Quatile	29087	317736	4052	3894.309	8.98	7572.18
IQR	22773	99360	2801	2064.844	7.36	4859.54
Upper Fence	63246.5	466776	8253.5	6991.575	20.02	14861.49
Lower Fence	-27845.5	69336	-2950.5	-1267.8	-9.42	-4576.67

As we can see red line the values out of lower fence and upper fence range. Cities Cheyenne, Gillette, are outliers.Since Gillette are pretty close to the upper fence and we only have 11 cities we should only remove Cheyenne.

Scatter Plot tools in alteryx : 2010Census Population vs. Total =Total Pawdacity Sales: