Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Key Decisions:

1. What decisions needs to be made?
   Recommend the city for Pawdacity's new store by analysing the relationship between historical data set and demographic data set,

2. What data is needed to inform those decisions?
   (1) monthly sales data for all Pawdacity stores in the year 2010
   (2) NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales
   (3) population numbers
   (4) demographic data

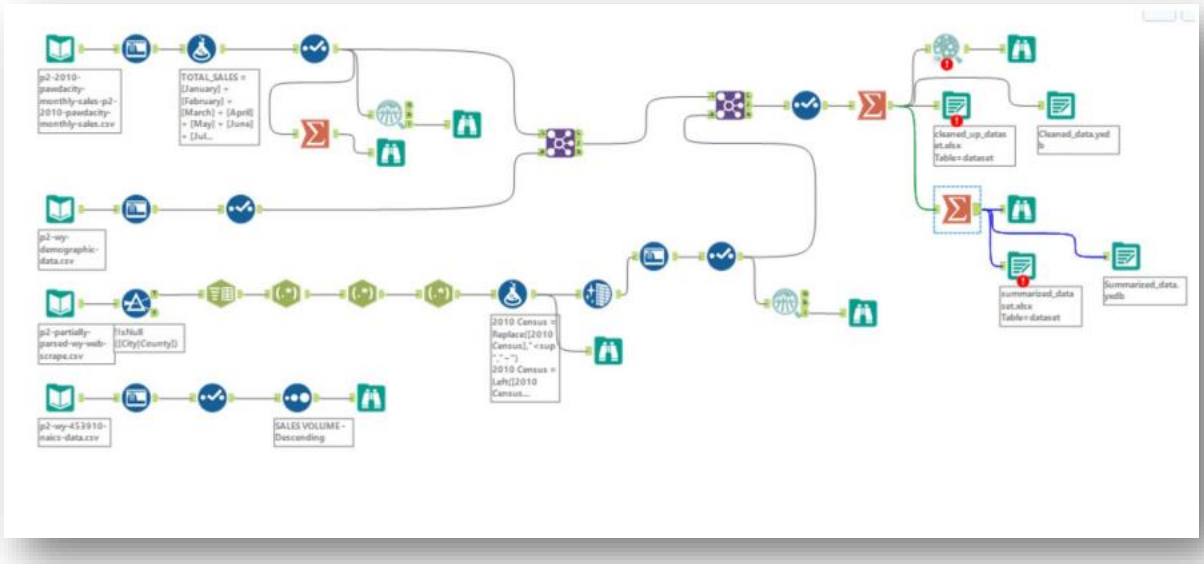
Step 2: Building the Training Set

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

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*

Alterix Flow:



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.

City	Total_Sales	2010 Census	County	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4585	Johnson	3115.5075	746	1.55	1819.5
Casper	317736	35316	Natrona	3894.3091	7788	11.16	8756.32
Cheyenne	917892	59466	Laramie	1500.1784	7158	20.34	14612.64
Cody	218376	9520	Park	2998.95696	1403	1.82	3515.62
Douglas	208008	6120	Converse	1829.4651	832	1.46	1744.08
Evanston	283824	12359	Uinta	999.4971	1486	4.95	2712.64
Gillette	543132	29087	Campbell	2748.8529	4052	5.8	7189.43
Powell	233928	6314	Park	2673.57455	1251	1.62	3134.18
Riverton	303264	10615	Fremont	4796.859815	2680	2.34	5556.49
Rock Springs	253584	23036	Sweetwater	6620.201916	4022	2.78	7572.18
Sheridan	308232	17444	Sheridan	1893.977048	2646	8.98	6039.71
Q1	226152	7917		1861.721074	1327	1.72	2923.41
Q3	312984	26061.5		3504.9083	4037	7.39	7380.805
IQR	86832	18144.5		1643.187226	2710	5.67	4457.395
Upper Fence	443232	53278.25		5969.689139	8102	15.895	14066.8975
Lower Fence	95904	-19299.75		-603.059765	-2738	-6.785	-3762.6825

After careful consideration:

1. **Cheyenne**: Cheyenne is also the capital city of Wyoming so a high population density, total population, and total families are expected. the scatter Plots demonstrate a linear relationship between total sales against each of those metrics. Therefore, Cheyenne is **kept**

2. **Gillette** : Because the total sales for this city is an outlier which can't be explained by other variables found in land area, population density & total families, Therefore, Gillette should be **removed** from the dataset.

3. **Rock Springs**: The city possesses a higher land area relative to other cities as shown in land area versus total sales yet the land area still fits the linear model with total sales. Thus it is **kept**

Scatterplot to find linear relationships: