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?

   Recommending the city for Pawdacity's newest store based on the predicted yearly sales.

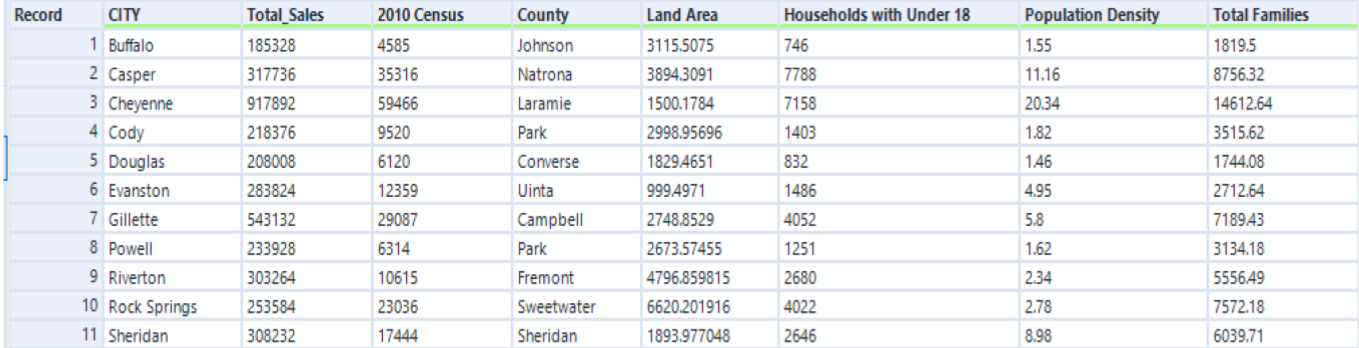
2. What data is needed to inform those decisions?

- Pawdacity stores monthly sales for 2010
- NAICS data of all competitor stores (for 12 months)
- A partially parsed data file for population numbers.
- Demographic data for of the state.

Step 2: Building the Training Set

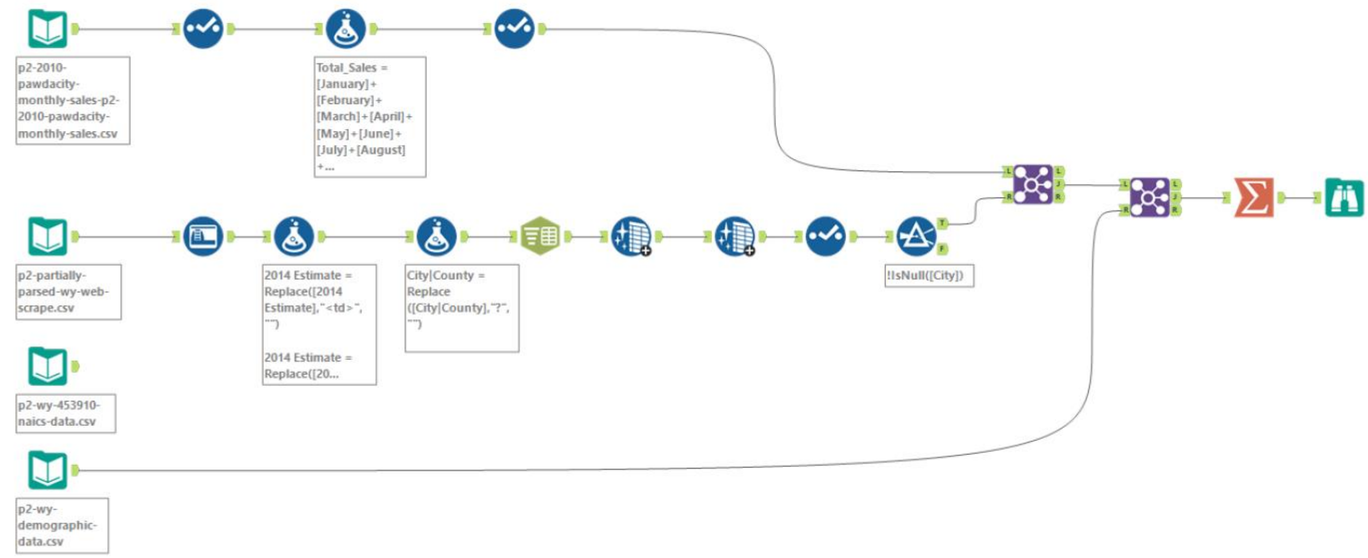
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

My results matched the given data.



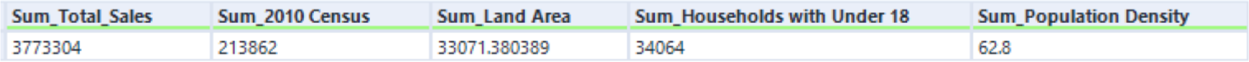
Record	CITY	Total_Sales	2010 Census	County	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185328	4585	Johnson	3115.5075	746	1.55	1819.5
2	Casper	317736	35316	Natrona	3894.3091	7788	11.16	8756.32
3	Cheyenne	917892	59466	Laramie	1500.1784	7158	20.34	14612.64
4	Cody	218376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

Workflow:


p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total_Sales = [January]+[February]+[March]+[April]+[May]+[June]+[July]+[August]+...
p2-partially-parsed-wy-web-scrape.csv
2014 Estimate = Replace([2014 Estimate],"<td>","")

2014 Estimate = Replace([20...
City|County = Replace([City|County],"?","")
!IsNull([City])
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv

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



Sum_Total_Sales	Sum_2010 Census	Sum_Land Area	Sum_Households with Under 18	Sum_Population Density
3773304	213862	33071.380389	34064	62.8

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*

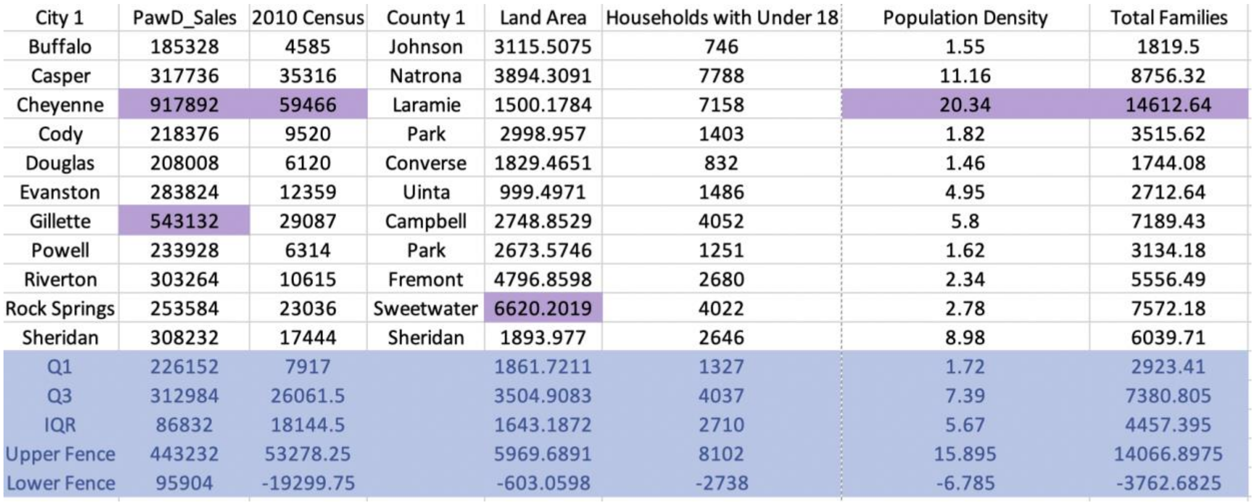
Step 3: Dealing with Outliers

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 used Excel function (QUARTILE.INC) to calculate Q1 and Q3 then,

- IQR = Q3 – Q1
- Upper Fence = Q3 + 1.5 * IQR
- Lower Fence = Q1 – 1.5 * IQR



City 1	PawD_Sales	2010 Census	County 1	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	2998.957	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	2673.5746	1251	1.62	3134.18
Riverton	303264	10615	Fremont	4796.8598	2680	2.34	5556.49
Rock Springs	253584	23036	Sweetwater	6620.2019	4022	2.78	7572.18
Sheridan	308232	17444	Sheridan	1893.977	2646	8.98	6039.71
Q1	226152	7917		1861.7211	1327	1.72	2923.41
Q3	312984	26061.5		3504.9083	4037	7.39	7380.805
IQR	86832	18144.5		1643.1872	2710	5.67	4457.395
Upper Fence	443232	53278.25		5969.6891	8102	15.895	14066.8975
Lower Fence	95904	-19299.75		-603.0598	-2738	-6.785	-3762.6825

Since we only have to remove one city, it is Cheyenne because the others are pretty close to the upper fence, they won't affect the data.

REF:

[1] https://github.com/kaishengteh/Predictive-Analytics-for-Business-Nanodegree/blob/master/2-Creating-an-Analytical-Dataset/2.1-Data-Cleanup.ipynb