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?

The most important decision that needs to be made is choosing the best city for Pawdacity's newest store, based on predicted yearly sales.

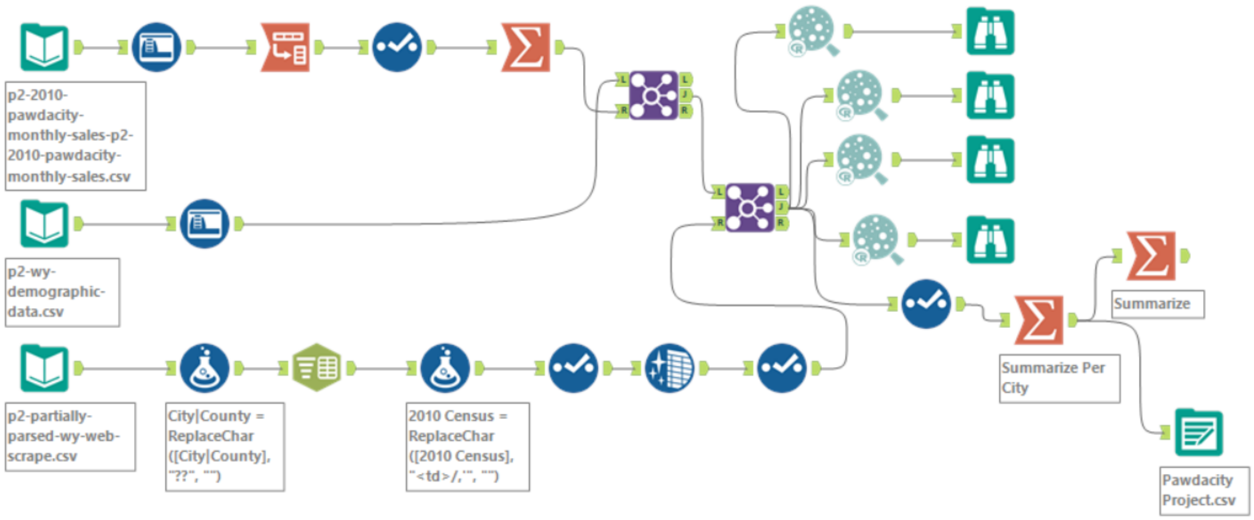
2. What data is needed to inform those decisions?

The data needed to inform this decision represents the monthly sales for all Pawdacity stores for the year of 2010, demographic data which will better reflect the location of these stores (land area, households with individuals under 18, population density). We look at these data at city level not at store level.

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

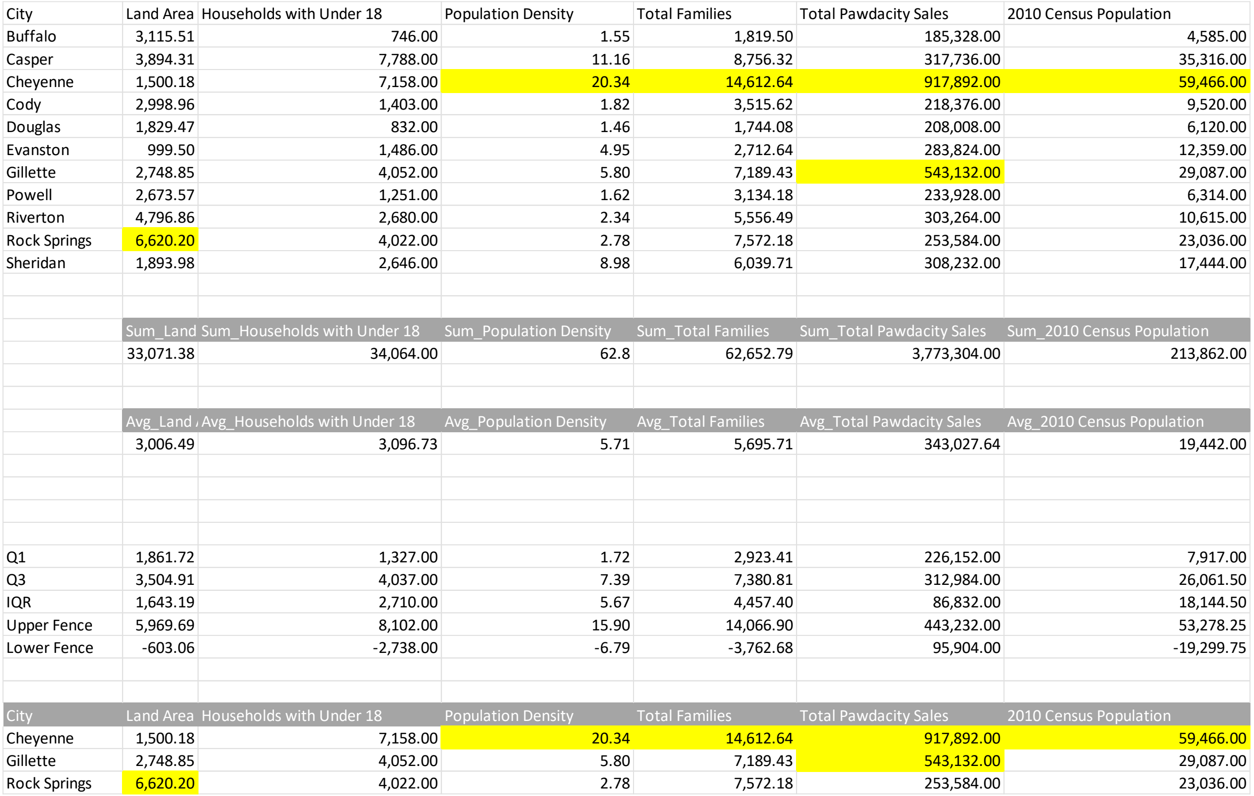

p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
City|County = ReplaceChar ([City|County], "??", "")
2010 Census = ReplaceChar ([2010 Census], "<td>/,'", "")
Summarize
Summarize Per City
Pawdacity Project.csv
L
J
R

Column	**Sum**	**Average**
Census Population	*213,862*	19,442.00
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.



City	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales	2010 Census Population
Buffalo	3,115.51	746.00	1.55	1,819.50	185,328.00	4,585.00
Casper	3,894.31	7,788.00	11.16	8,756.32	317,736.00	35,316.00
Cheyenne	1,500.18	7,158.00	20.34	14,612.64	917,892.00	59,466.00
Cody	2,998.96	1,403.00	1.82	3,515.62	218,376.00	9,520.00
Douglas	1,829.47	832.00	1.46	1,744.08	208,008.00	6,120.00
Evanston	999.50	1,486.00	4.95	2,712.64	283,824.00	12,359.00
Gillette	2,748.85	4,052.00	5.80	7,189.43	543,132.00	29,087.00
Powell	2,673.57	1,251.00	1.62	3,134.18	233,928.00	6,314.00
Riverton	4,796.86	2,680.00	2.34	5,556.49	303,264.00	10,615.00
Rock Springs	6,620.20	4,022.00	2.78	7,572.18	253,584.00	23,036.00
Sheridan	1,893.98	2,646.00	8.98	6,039.71	308,232.00	17,444.00

	Sum_Land	Sum_Households with Under 18	Sum_Population Density	Sum_Total Families	Sum_Total Pawdacity Sales	Sum_2010 Census Population
	33,071.38	34,064.00	62.8	62,652.79	3,773,304.00	213,862.00

	Avg_Land	Avg_Households with Under 18	Avg_Population Density	Avg_Total Families	Avg_Total Pawdacity Sales	Avg_2010 Census Population
	3,006.49	3,096.73	5.71	5,695.71	343,027.64	19,442.00

Q1	1,861.72	1,327.00	1.72	2,923.41	226,152.00	7,917.00
Q3	3,504.91	4,037.00	7.39	7,380.81	312,984.00	26,061.50
IQR	1,643.19	2,710.00	5.67	4,457.40	86,832.00	18,144.50
Upper Fence	5,969.69	8,102.00	15.90	14,066.90	443,232.00	53,278.25
Lower Fence	-603.06	-2,738.00	-6.79	-3,762.68	95,904.00	-19,299.75

City	Land Area	Households with Under 18	Population Density	Total Families	Total Pawdacity Sales	2010 Census Population
Cheyenne	1,500.18	7,158.00	20.34	14,612.64	917,892.00	59,466.00
Gillette	2,748.85	4,052.00	5.80	7,189.43	543,132.00	29,087.00
Rock Springs	6,620.20	4,022.00	2.78	7,572.18	253,584.00	23,036.00

After creating the dataset, I found that there are cities with outliers by using the IQR method in Excel. As seen in the above table, there are three outliers: Cheyenne, Gillette and Rock Springs. I have decided to remove Cheyenne from my dataset because there are too many values that deviate, and this would definitely have an impact on my predictions. Also, these deviations are in the key variables, which most likely will have an impact on our future predictions.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.