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 needs a recommendation for the city to set their newest store.

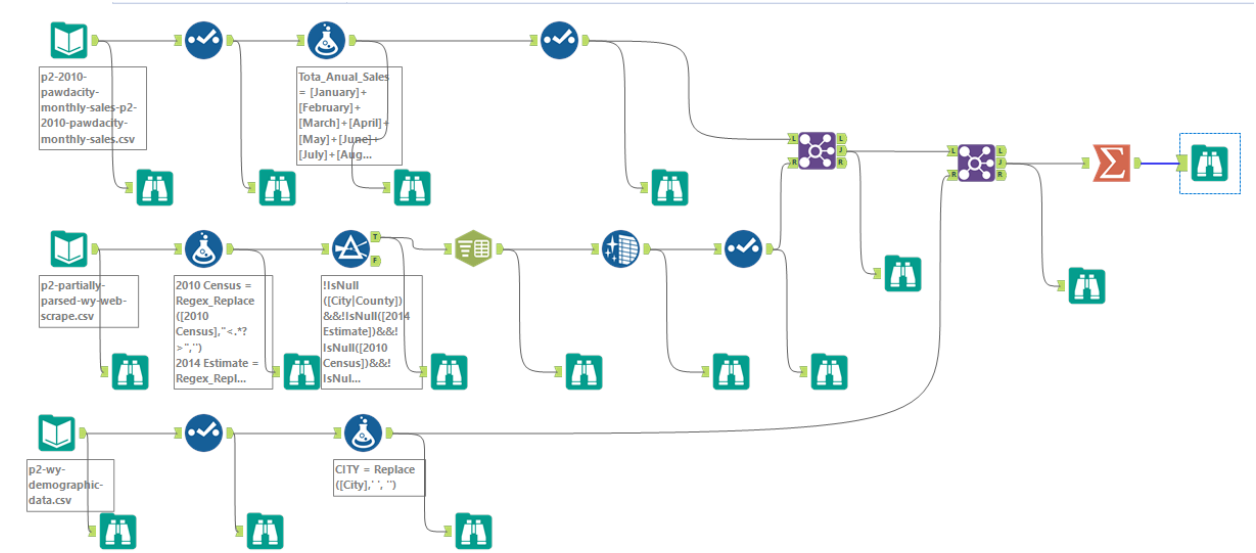
2. What data is needed to inform those decisions?

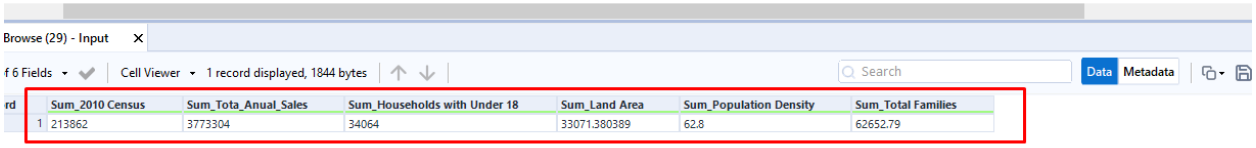
In order to analyze and give a recommendation for this decision, it is needed:

- The Pawdacity stores sales per year per city
- Sale volumes of competitors
- Land area,
- Population density,
- Total families
- Households with kids under 18, (families are more likely to have pets).

Step 2: Building the Training Set

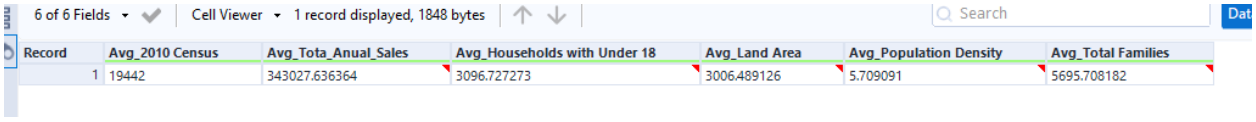
Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.





	Sum_2010 Census	Sum_Tota_Anual_Sales	Sum_Households with Under 18	Sum_Land Area	Sum_Population Density	Sum_Total Families
1	213862	3773304	34064	33071.380389	62.8	62652.79

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



Record	Avg_2010 Census	Avg_Tota_Anual_Sales	Avg_Households with Under 18	Avg_Land Area	Avg_Population Density	Avg_Total Families
1	19442	343027.636364	3096.727273	3006.489126	5.709091	5695.708182

Column	Sum	Average
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*343,027.63*
Households with Under 18	*34,064*	*3,096.72*
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.

After getting the Lower Fence and Upper Fence, it is possible to observe a few outliers such as Cheyenne, Gillete, and RockSprings that may be convenient to remove, but considering that for the instructions for this project is to only remove one, I would choose Chayanne, this city has most of their values out of upper fence, as it is possible to see in the table below, where every value that is out of the Lowe/Upper Fence is highlighted in orange.

CITY	Tota_Anual_ Sales	2010 Census	Land Area	Households with Under 18	Population Density	Total Families	
Cheyenne	917,892.00	59,466.00	1,500.18	7,158.00	20.34	14,612.64	OUTLIER
Gillette	543,132.00	29,087.00	2,748.85	4,052.00	5.8	7,189.43	OUTLIER
Casper	317,736.00	35,316.00	3,894.31	7,788.00	11.16	8,756.32	
Sheridan	308,232.00	17,444.00	1,893.98	2,646.00	8.98	6,039.71	
Riverton	303,264.00	10,615.00	4,796.86	2,680.00	2.34	5,556.49	
Evanston	283,824.00	12,359.00	999.5	1,486.00	4.95	2,712.64	
RockSprings	253,584.00	23,036.00	6,620.20	4,022.00	2.78	7,572.18	OUTLIER
Powell	233,928.00	6,314.00	2,673.57	1,251.00	1.62	3,134.18	
Cody	218,376.00	9,520.00	2,998.96	1,403.00	1.82	3,515.62	
Douglas	208,008.00	6,120.00	1,829.47	832	1.46	1,744.08	
Buffalo	185,328.00	4,585.00	3,115.51	746	1.55	1,819.50	
Q1	226,152.00	7,917.00	1,861.72	1,327.00	1.72	2,923.41	
Q3	312,984.00	26,061.50	3,504.91	4,037.00	7.39	7,380.81	
InterQuartile(Q3-Q1)	86,832.00	18,144.50	1,643.19	2,710.00	5.67	4,457.40	
Lower Fence (Q1-1.5IQR)	95,904.00	-19,299.75	-603.06	-2,738.00	-6.79	-3,762.68	
Upper Fence (Q1+1.5IQR)	443,232.00	53,278.25	5,969.69	8,102.00	15.9	14,066.90	

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.