Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**

Choose a city for a 14th Pawdacity store.

2. **What data is needed to inform those decisions?**

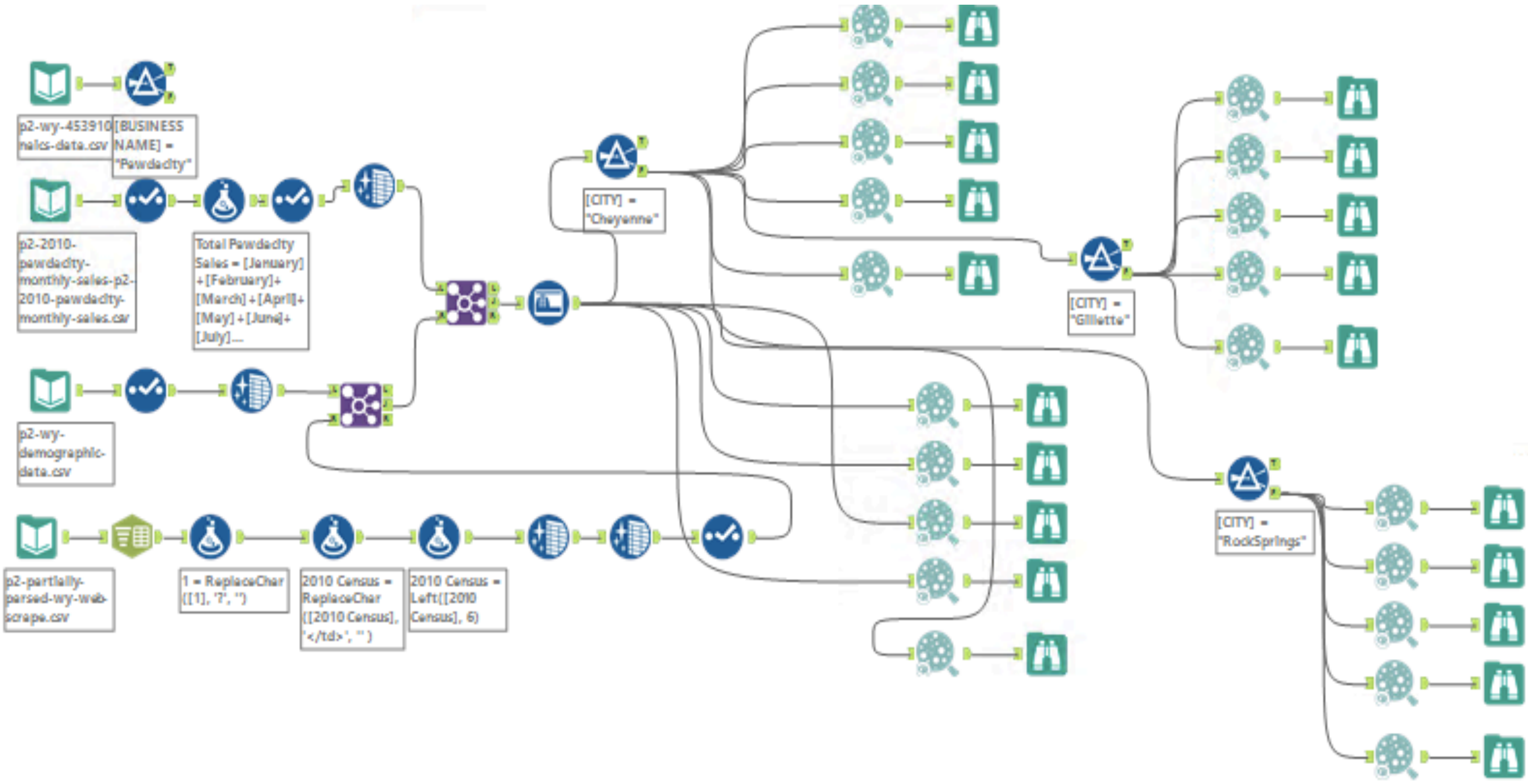
Predicted yearly sales data to choose a city with highest predicted yearly sales for a new store.

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.

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*

Please, find my workflow below.


p2-wy-453910-naics-data.csv
[BUSINESS NAME] = "Pawdacity"
p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total Pawdacity Sales = [January] +[February]+ [March] +[April]+ [May] +[June]+ [July]...
p2-wy-demographic-data.csv
p2-partially-parsed-wy-web-scrape.csv
1 = ReplaceChar ([1], '?', '')
2010 Census = ReplaceChar ([2010 Census], '</td>', '')
2010 Census = Left([2010 Census], 6)
[CITY] = "Cheyenne"
[CITY] = "Gillette"
[CITY] = "RockSprings"

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.

I have highlighted outliners in the orange colour. Clearly seen that Cheyenne has outliners in 4 categories (Total Pawdacity Sales, Population Density, Total Families, 2010 Census), but they seem logically dependent on each other. I suggest to exclude Gilette as this city doesn't have logical explanation for that outlier value.

CITY	Total Pawdacity Sales	Land Area	Households with Under 18	Population Density	Total Families	2010 Census
Buffalo	185328	3115.5075	746	1.55	1819.5	4,585
Casper	317736	3894.3091	7788	11.16	8756.32	35,316
Cheyenne	917892	1500.1784	7158	20.34	14612.64	59,466
Cody	218376	2998.95696	1403	1.82	3515.62	9,520
Douglas	208008	1829.4651	832	1.46	1744.08	6,120
Evanston	283824	999.4971	1486	4.95	2712.64	12,359
Gillette	543132	2748.8529	4052	5.8	7189.43	29,087
Powell	233928	2673.57455	1251	1.62	3134.18	6,314
Riverton	303264	4796.859815	2680	2.34	5556.49	10,615
RockSprings	253584	6620.201916	4022	2.78	7572.18	23,036
Sheridan	308232	1893.977048	2646	8.98	6039.71	17,444
SUM	3,773,304	33,071	34,064	63	62,653	213,862
Average	343,027.64	3,006.49	3,096.73	5.71	5,695.71	19,442.00

Total Pawdacity Sales	
Q1	226152
Q3	312984
IQR	86832
Lower fence	95904
Upper fence	443232
Outliers_Count	2
Outliers	Cheyenne
	Gillette

Land Area	
Q1	1861.721074
Q3	3504.9083
IQR	1643.187226
Lower fence	-603.059765
Upper fence	5969.689139
Outliers_Count	1
Outliers	RockSprings

Households with Under 18	
Q1	1327
Q3	4037
IQR	2710
Lower fence	-2738
Upper fence	8102
Outliers_Count	0
Outliers	

Population Density	
Q1	1.72
Q3	7.39
IQR	5.67
Lower fence	-6.785
Upper fence	15.895
Outliers_Count	1
Outliers	Cheyenne

Total Families	
Q1	2923.41
Q3	7380.805
IQR	4457.395
Lower fence	-3762.6825
Upper fence	14066.8975
Outliers_Count	1
Outliers	Cheyenne

2010 Census	
Q1	7917
Q3	26061.5
IQR	18144.5
Lower fence	-19299.75
Upper fence	53278.25
Outliers_Count	1
Outliers	Cheyenne