Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

Key Decisions:

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

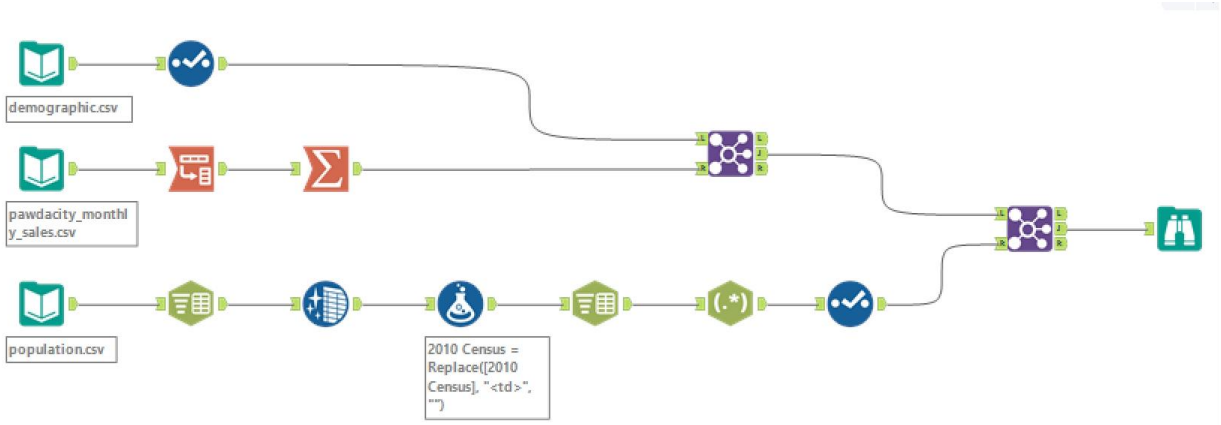
Which city is the best for opening a 14th store?

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

- The monthly sales data for all of the Pawdacity stores for the year 2010.
- NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
- A partially parsed data file that can be used for population numbers.
- Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyomin.

Step 2: Building the Training Set

This is the workflow performed in Alteryx.



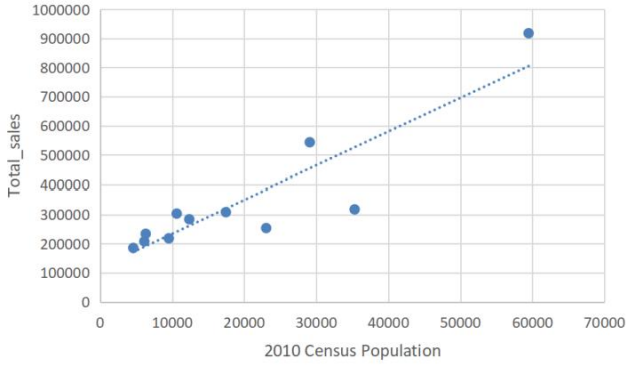
Column	Sum	Average
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

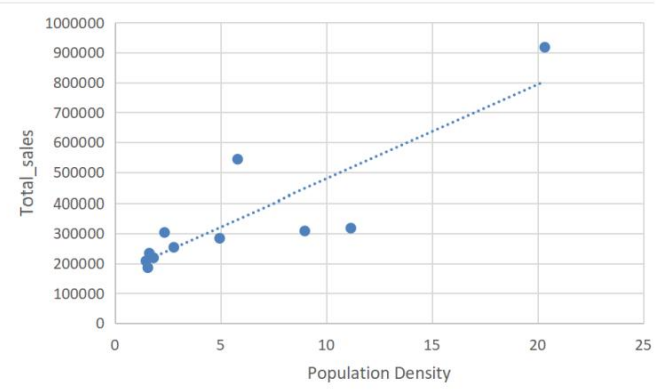
Step 3: Dealing with Outliers

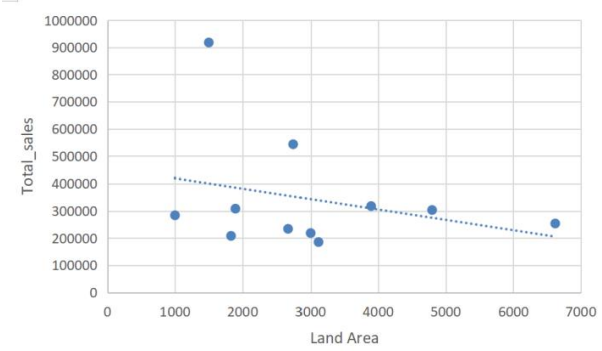
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

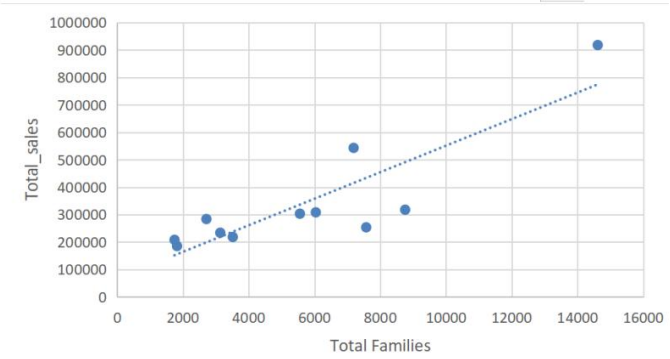
City	Land Area	Households with Under 18	Population Density	Total Families	2010 Census Population	Total_sales
Evanston	999.4970703	1486	4.949999809	2712.639893	12359	283824
Douglas	1829.465088	832	1.460000038	1744.079956	6120	208008
Casper	3894.309082	7788	11.15999985	8756.320313	35316	317736
Rock Spring	6620.202148	4022	2.779999971	7572.180176	23036	253584
Cheyenne	1500.178345	7158	20.34000015	14612.63965	59466	917892
Cody	2998.957031	1403	1.820000052	3515.620117	9520	218376
Gillette	2748.852783	4052	5.800000191	7189.430176	29087	543132
Riverton	4796.859863	2680	2.339999914	5556.490234	10615	303264
Powell	2673.574463	1251	1.620000005	3134.179932	6314	233928
Buffalo	3115.507568	746	1.549999952	1819.5	4585	185328
Sheridan	1893.977051	2646	8.979999542	6039.709961	17444	308232
Q1	1861.721069	1327	1.720000029	2923.409912	7917	226152
Q3	3504.908325	4037	7.389999866	7380.805176	26061.5	312984
IQR	1643.187256	2710	5.669999838	4457.395264	18144.5	86832
Upper fence	5969.689209	8102	15.89499962	14066.89807	53278.25	443232
Lower fence	-603.0598145	-2738	-6.784999728	-3762.682983	-19299.75	95904

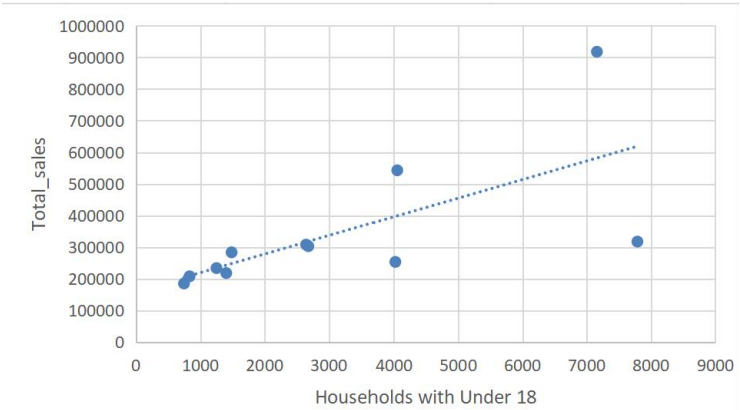
We can see that values marked red with pink background are outliers in each variables, then find matching cities (Rock Springs, Cheyenne, and Gillette) that were outliers in the training set. Using scatter plot to investigate the relationship between each predictor with the variable “Total_sales” as below:











We can see that Cheyenne has outliers in most of variables. Doing a small research and I find that Cheyenne is the capital and most populous city of the U.S. state of Wyoming. It makes sense for Cheyenne to be different from other cities in Wyoming. Although Cheyenne has second smallest land area, this city has the largest total sales.

Therefore, I will remove the city Cheyenne from the training set to avoid biased model due to the effect of outliers. See the scatter plot showing the relationship between land area and total sales below, now the slope of the line is increased after removing outlier.