Project 2.1: Data Cleanup

Step 1: Business and Data Understanding

1. What decisions needs to be made?

To decide in which city Pawdacity should expand and open its 14th store based on the predicted annual sales for the given cities in Wyoming.

2. What data is needed to inform those decisions?

A couple of files are provided for this case and they'll be utilized as follow:

- *p2-2010-pawdacity-monthly-sales.csv* – the historical annual sales will be used in the prediction model.
- *p2-partially-parsed-wy-web-scrape.csv* –this will be used to extract data about the population in each city.
- *p2-wy-demographic-data.csv* - this file summarized the demographic data for each city and county in Wyoming.

However, another set of data *"p2-wy-453910-naics-data.csv "* has been given, but that will be used later in the second part of this project.

Step 2: Building the Training Set

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*

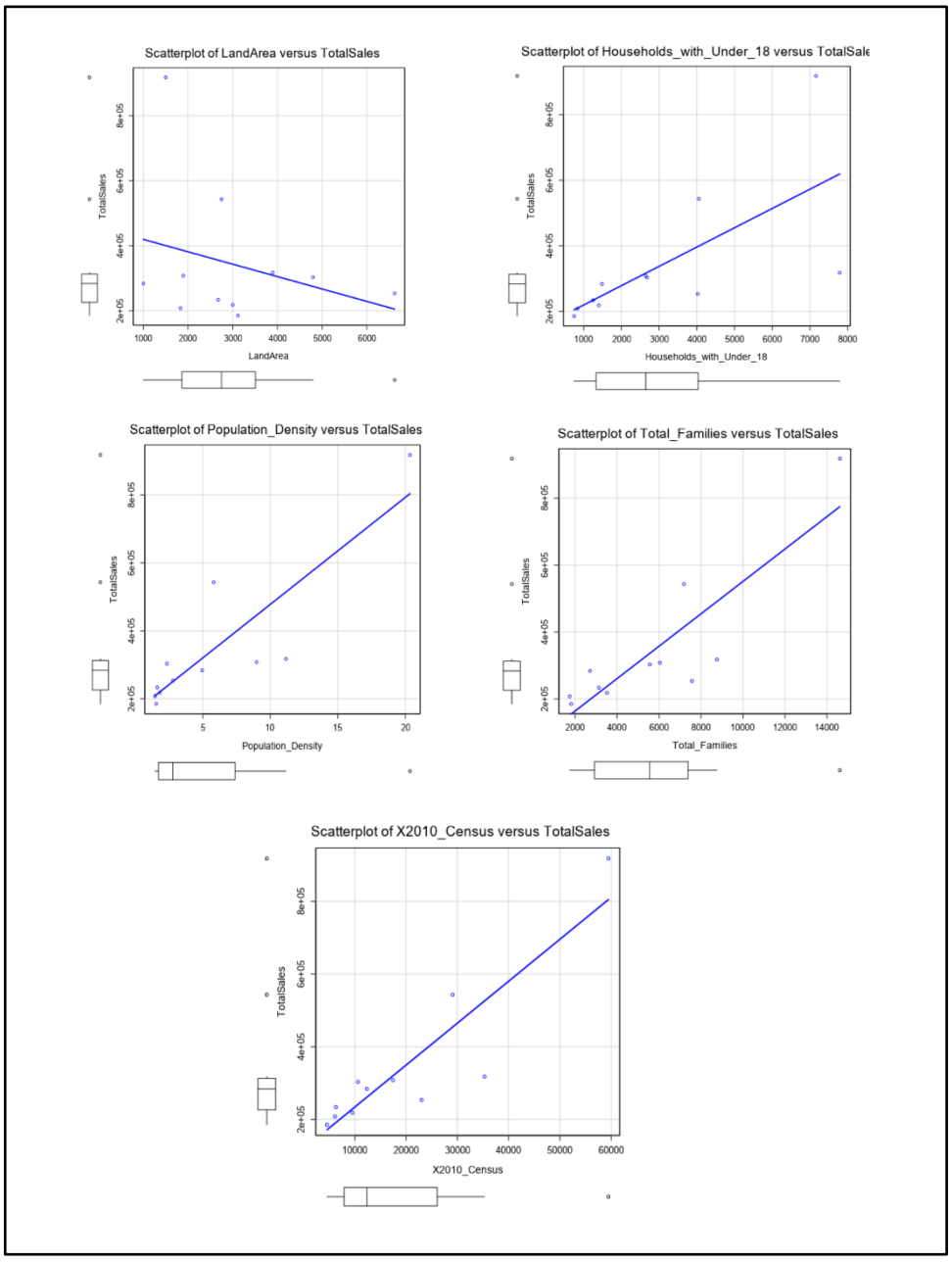
Step 3: Dealing with Outliers

To detect the outliers, a scatterplot has been generated in order to illustrate the relation between the Pawdacity total sales and other variables. Also, the IQR range and the upper fence have been calculated to help identifying the outliers -all the graphs and details are provided in the next page-.
Based on that, the following cities need to be investigated for the values below;

- Cheyenne City for Total Sale, Census Population, Total Families and Population Density
- Rock Springs for Land Area
- Gillette for Total sales.

Since we're analyzing the data on the city level, I can consider having high sales in Cheyenne is justifiable compared to its high papulation. However, Rock Springs is having only one outlier "Land Area" and that follows the same downward pattern of the line. Lastly, Gillette is having high sales while

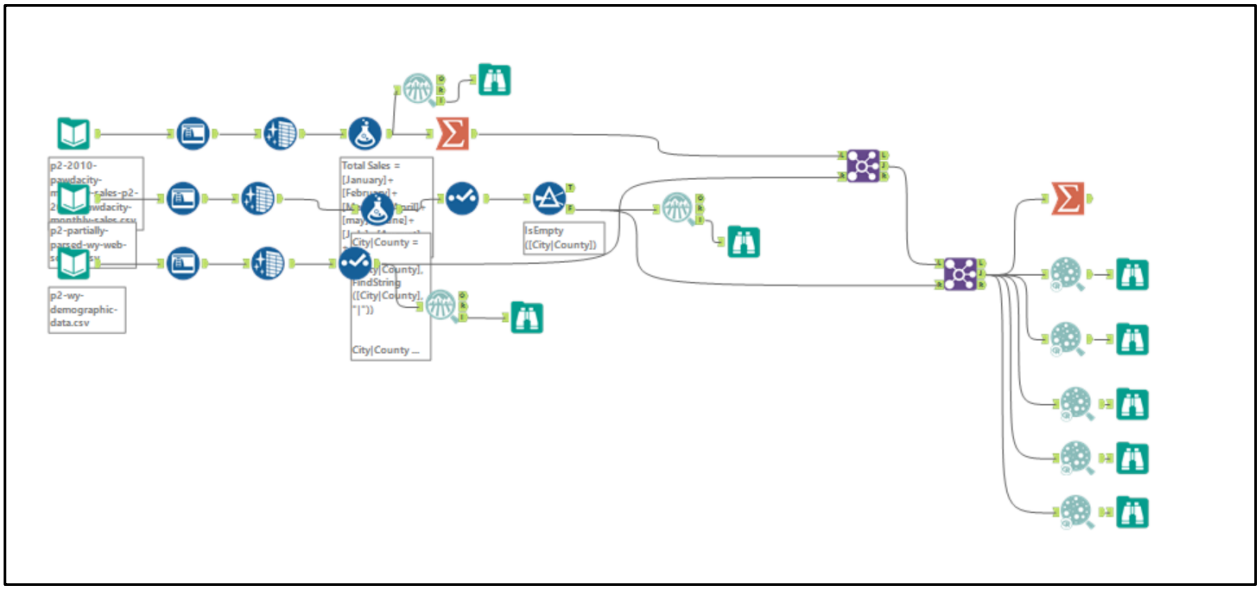
still maintaining other values within the IQR range. Hence, I recommend removing Gillette and keeping the other two cities as their values can be considered appropriate and justifiable unlike the inordinate value of sales in Gillette.



Graph 1: Scatterplot of TotalSale & other variables

City	TotalSales	LandArea	Households with Under 18	Population Density	Total Families	2010 Census
Buffalo	185,328.00	3115.51	746	1.55	1819.50	4585.00
Casper	317,736.00	3894.31	7788	11.16	8756.32	35316.00
Cheyenne	917,892.00	1500.18	7158	20.34	14612.64	59466.00
Cody	218,376.00	2998.96	1403	1.82	3515.62	9520.00
Douglas	208,008.00	1829.47	832	1.46	1744.08	6120.00
Evanston	283,824.00	999.50	1486	4.95	2712.64	12359.00
Gillette	543,132.00	2748.85	4052	5.80	7189.43	29087.00
Powell	233,928.00	2673.57	1251	1.62	3134.18	6314.00
Riverton	303,264.00	4796.86	2680	2.34	5556.49	10615.00
RockSprings	253,584.00	6620.20	4022	2.78	7572.18	23036.00
Sheridan	308,232.00	1893.98	2646	8.98	6039.71	17444.00
Q1	226,152.00	1861.72	1327	1.72	2923.41	7917
Q2	283,824.00	2748.85	2646	2.78	5556.49	12359
Q3	312,984.00	3504.91	4037	7.39	7380.805	26061.5
IQR	86,832.00	1643.19	2710	5.67	4457.395	18144.5
Upper Fence	443,232.00	5969.69	8102	15.895	14066.8975	53278.25
Lower Fence	95,904.00	-128386.28	-2738	-6.785	-3762.6825	-19299.75

Table 1: IQR Calculation



Graph 2: Alteryx Workflow