Project 2.1: Data Cleanup

https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   Well, From the available data we need to provide information to the decision maker in Pawdacity (a leading pet store chain in Wyoming). The analysis is to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

2. What data is needed to inform those decisions?

   Because the monthly sales represent the year of 2010, the following data are needed:

   1- City – The available city that Pawdacity chain works in.
   2- 2010 Census Population
   3- Total Pawdacity Sales for 2010 – Aggregated from the monthly sales.
   4- Household with under 18 -
   5- Land Area
   6- Population Density
   7- Total Families

Step 2: Building the Training Set

Column	Sum	Average
Census Population	213,862	19,442
Total Pawdacity Sales	3,773,304	343,028
Households with Under 18	34,064	3,097
Land Area	33,071	3,006
Population Density	63	5.7
Total Families	62,653	5,696

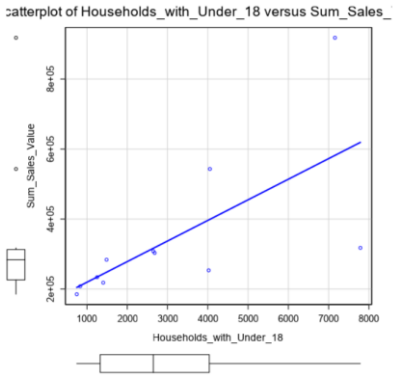
Note: Screen shot for the result (By Alteryx):

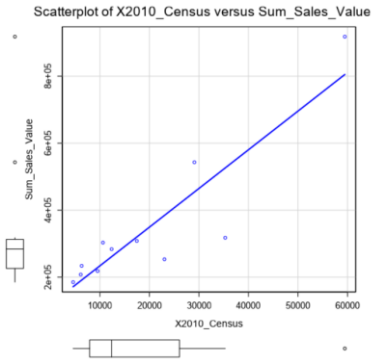
Record	Avg_2010 Census	Avg_Sum_Sales Value	Avg_Households with Under 18	Avg_Land Area	Avg_Population Density	Avg_Total Families
1	19442	343027.636364	3096.727273	3006.489126	5.709091	5695.708182

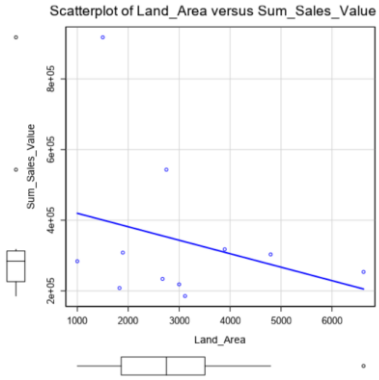
Step 3: Dealing with Outliers

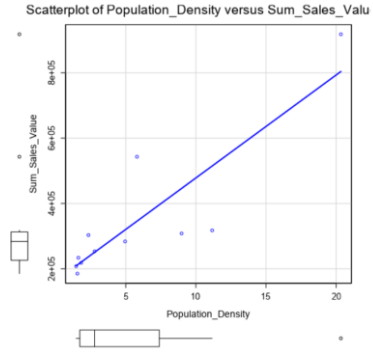
A- Charts To help deciding what outlier to remove:

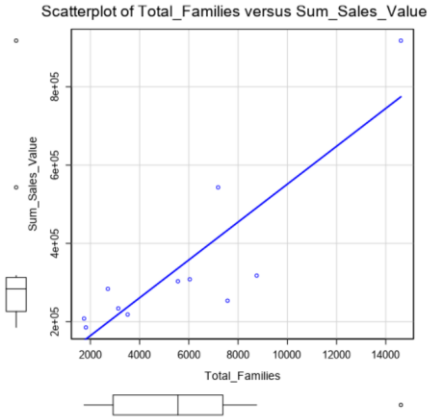
1- I will show below the scatter plots for all predictors (By Alteryx):











2- I will provide information about (IQR, Upper Fence, and Lower Fence) by Excel:

Field	Q1	Q3	IQR	Upper fence	Lower fence
2010 Census	7,917.00	26,061.50	18,144.50	53,278.25	(19,299.75)
Land Area	1,861.72	3,504.91	1,643.19	5,969.69	(603.06)
Households with Under 18	1,327.00	4,037.00	2,710.00	8,102.00	(2,738.00)
Population Density	1.72	7.39	5.67	15.90	(6.79)
Total Families	2,923.41	7,380.81	4,457.40	14,066.90	(3,762.68)
Total Sales Value	226,152.00	312,984.00	86,832.00	443,232.00	95,904.00

3- Histogram For all the predictors (By Alteryx):

4- Summary of the final Dataset (By Alteryx):

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	2010 Census	Numeric	4585	59466	12359	16616.018584	0	11	19442
2	Households with Under 18	Numeric	746	7788	2646	2453.003061	0	11	3096.727273
3	Land Area	Numeric	999.4971	6620.201916	2748.8529	1617.460342	0	11	3006.489126
4	Population Density	Numeric	1.46	20.34	2.78	5.849685	0	11	5.709091
5	Sum_Sales Value	Numeric	185328	917892	283824	213538.712215	0	11	343027.636364
6	Total Families	Numeric	1744.08	14612.64	5556.49	3816.04966	0	11	5695.708182

5- Cities that are outliers (Above the upper Fence):

Cheyenne - 2010 Census
Rock Springs - Land Area
Cheyenne - Population Density
Cheyenne - Total Families
Gillette and Cheyenne - Total Sales Value

B-Analysis and Decision:

From A.5 Cities that are outliers (Above the upper Fence), I need to decide among them which city the dataset analysis will allow me to remove. The cities are:

- **Cheyenne.**
- **Rock Springs.**
- **Gillette.**

Cheyenne:

Considered outlier in total sales, total families, population density, and census of 2010. And that reasonable to be outlier but I cannot remove it because the correlation between the three factors is very high. Please see the following image:

Record	FieldName	Sum_Sales Value	2010 Census	Land Area	Households with Under 18	Population Density	Total Families
1	Sum_Sales Value	1	0.898099	-0.288898	0.676012	0.862894	0.86466
2	2010 Census	0.898099	1	-0.061587	0.911883	0.927702	0.968005
3	Land Area	-0.288898	-0.061587	1	0.180704	-0.317244	0.099389
4	Households with Under 18	0.676012	0.911883	0.180704	1	0.815756	0.907242
5	Population Density	0.862894	0.927702	-0.317244	0.815756	1	0.884792
6	Total Families	0.86466	0.968005	0.099389	0.907242	0.884792	1

Rock Springs:

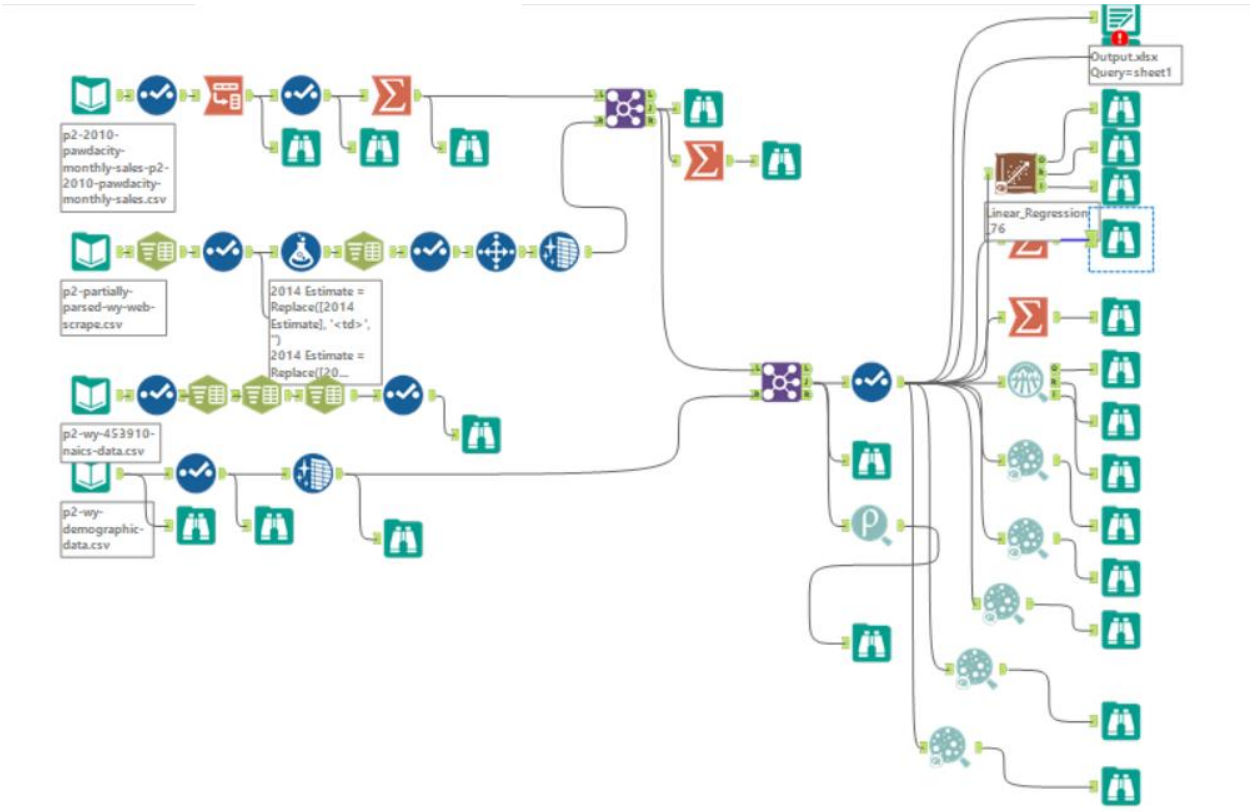
It is a city considered outlier in the Land Area. Its situation will affect the decision that management will make. Because it matches the request of the management about the city that needs to put two more stores in it. The sales in the city almost the average among all the cities.

Gillette:

The sales value is very high, but the Land Area is below the average. So, this information cannot be explained. **Gillette** will be the outlier I can remove. It is not like the case of **Cheyenne** and not the same case of **Rock Springs**, the main reason to remove this city, as follow:

- Its case to be an outlier cannot be explained and need more investigation.
- No need to increase the number of the stores.
- The sales value does not match with the correlation table.

C- Workflow (By Alteryx):



D-Tools I used in Alteryx: