CREATE AN ANALYTICAL DATASET: PAWDACITY

Business and Data Understanding

Pawdacity would like to expand and open a 14th Store.

Key Decisions:

Pawdacity is a leading pet store chain in Wyoming currently with 13 stores throughout the state, to expand and open a 14th store. To make the recommendation city for the new store, we need make an analysis from previous years sales of each city.

To be able to make an informed decision we need to have the following:

- Sales data for of the Pawdacity stores of 2010
- Data on the population records for each city.
- Demographic data of each city in the state of Wyoming.

Data that is needed to make the decision is the following:

City
2010 Census Population
Total Pawdacity Sales
Households with Under 18
Land Area
Population Density
Total Families

Building the Training Set

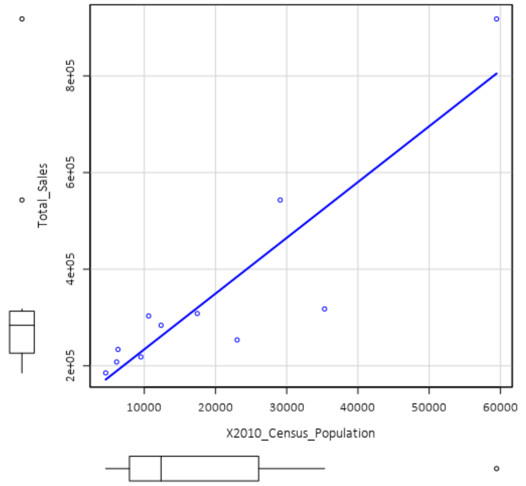
Here we focus on cleaning up the data set, and blend on city level and not at store level. Data provided is only city wide, so any analysis at store level would not be sufficient.

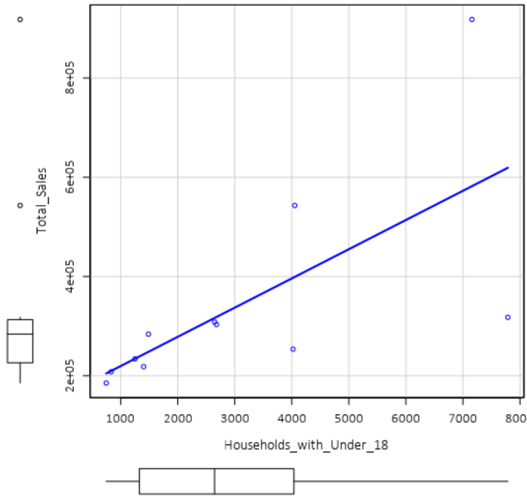
Data Validation

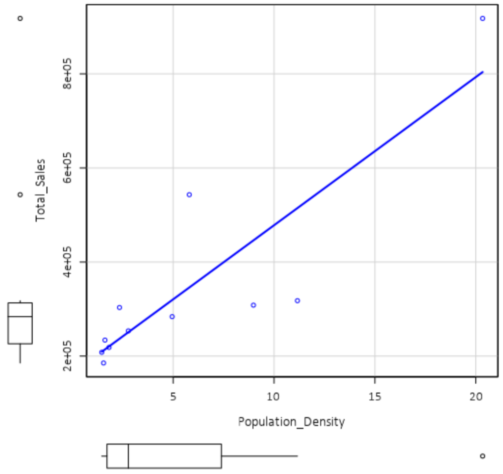
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3096.73*
Land Area	*33,071*	*3096.73*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

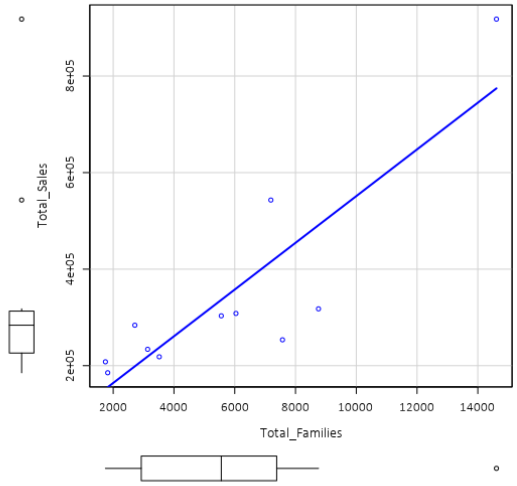
Dealing with Outliers

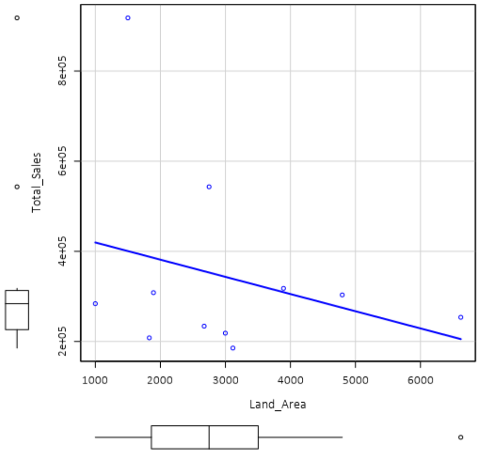
Scatter Plots to help visualize each predictor variable against the total city sales for Pawdacity.


Scatterplot of X2010_Census_Population versus Total_Sales


Scatterplot of Households_with_Under_18 versus Total_Sales


Scatterplot of Population_Density versus Total_Sales


Scatterplot of Total_Families versus Total_Sales


Scatterplot of Land_Area versus Total_Sales

City	Total_Sales	2010_Census Population	Land Area	Households with Under 18	Population Density	Total Families
Buffalo	185328	4,585	3,115.51	746	1.55	1,819.50
Casper	317736	35,316	3,894.31	7,788	11.16	8,756.32
Cheyenne	917892	59,466	1,500.18	7,158	20.34	14,612.64
Cody	218376	9,520	2,998.96	1,403	1.82	3,515.62
Douglas	208008	6,120	1,829.47	832	1.46	1,744.08
Evanston	283824	12,359	999.50	1,486	4.95	2,712.64
Gillette	543132	29,087	2,748.85	4,052	5.80	7,189.43
Powell	233928	6,314	2,673.57	1,251	1.62	3,134.18
Riverton	303264	10,615	4,796.86	2,680	2.34	5,556.49
Rock Springs	253584	23,036	6,620.20	4,022	2.78	7,572.18
Sheridan	308232	17,444	1,893.98	2,646	8.98	6,039.71
Q1	**226152**	**7917**	**1861.721074**	**1327**	**1.72**	**2923.41**
Q3	**312984**	**26061.5**	**3504.9083**	**4037**	**7.39**	**7380.805**
IQR	**86832**	**18144.5**	**1643.187226**	**2710**	**5.67**	**4457.395**
Upper Fence	**443232**	**53278.25**	**5969.689139**	**8102**	**15.895**	**14066.8975**
Lower Fence	**95904**	**-19299.75**	**-603.059765**	**-2738**	**-6.785**	**-3762.6825**

Outliers

Calculation for the quartiles Q1 and Q3 using Excel, ‘QUARTILE.INC’
Interquartile Range: IQR Q3 – Q1
Upper Fence: Q3 + 1.5 IQR
Lower Fence: Q1 – 1.5 IQR

There are 3 Outliers in the blended dataset: Cheyenne, Gillette, Rock Springs, these are defined as the column values are either above the Upper Fence, or below the Lower Fence.

Cheyenne and Gillette do flag as an outlier, based on Total Sales.

Cheyenne also is flagged as an outlier based on population.

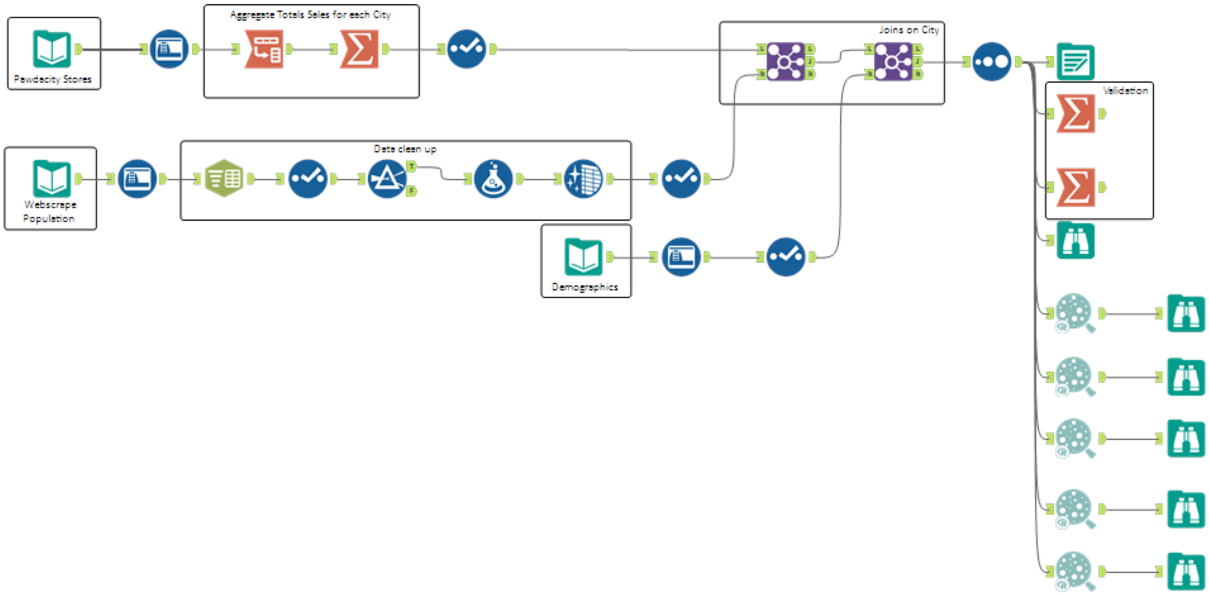
Rock Springs is an outlier due to have most land area.

Keeping Cheyenne in the data set would allow us to run a model against more populated cities and keeping Rock Springs will account for cities that have a bigger land area.

When we look at Gillette a little closer, we see that although it has high total sales, its population is almost 20,000 lower than Cheyenne. Also, the Total number of Families in Gillette is almost half of that of Cheyenne.

Therefore, we should drop Gillette from the Dataset.

Alteryx Workflow



Resources

https://knowledge.udacity.com/questions/166424