Data Cleaning

By Moamen Mohasseb

The Business Problem

Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Step 1: Business and Data Understanding

1. which City to open a new store?
   Factors to choose specific city like population density, total families …etc.
2. What is the predicted yearly sales for this city?
   Factors to predict city with highest yearly sales like other stores sales , population density … etc.

Step 2: Building the Training Set

We consolidating the data at the city level sum and average

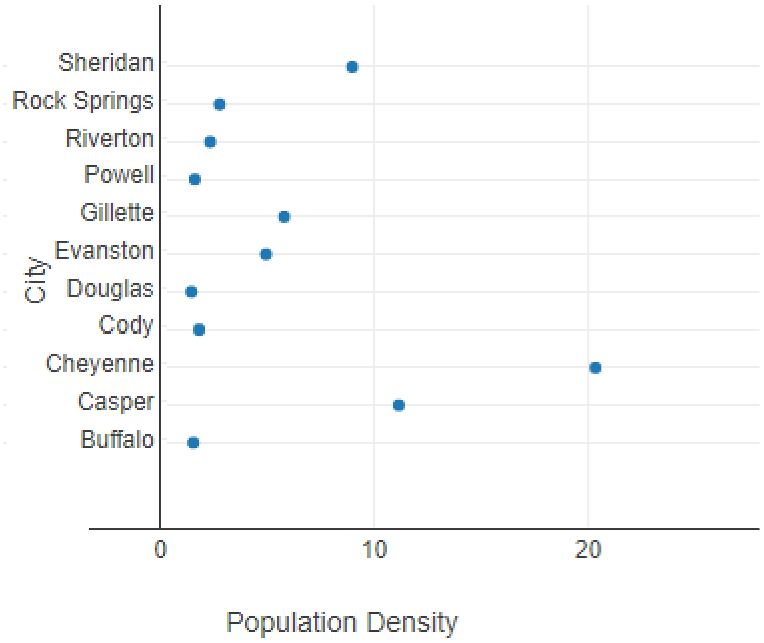
column	sum	average
Total Pawdacity Sales	3773304.00	343027.64
Population Density	62.80	5.71
Total Families	62652.79	5695.71
Households with Under 18	34064.00	3096.73
Land Area	33071.38	3006.49
Census Population	213862.00	19442.00

Step 3: Dealing with Outlier

There are two City with outliers in the training set : Cheyenne and Gilette. I have decided to remove Cheyenne because it has outliers in “Total Pawdacity Sales” due to large “Population Density” as we can see in the below chart .for “Gillette” I will keep it by reduce “Total Pawdacity Sales” to the upper fence 466776 .

City	County	Land Area	Households with Under 18	Population Density	Total Families	2014 Estimate	2010 Census	2000 Census	Total Sales
Buffalo	Johnson	3115.51	746.00	1.55	1819.50	4615.00	4.00	4585.00	185328.00
Casper	Natrona	3894.31	7788.00	11.16	8756.32	40086.00	35.00	35316.00	317736.00
Cheyenne	Laramie	1500.18	7158.00	20.34	14612.64	62845.00	59.00	59466.00	917892.00
Cody	Park	2998.96	1403.00	1.82	3515.62	9740.00	9.00	9520.00	218376.00
Douglas	Converse	1829.47	832.00	1.46	1744.08	6423.00	6.00	6120.00	208008.00
Evanston	Uinta	999.50	1486.00	4.95	2712.64	12190.00	12.00	12359.00	283824.00
Gillette	Campbell	2748.85	4052.00	5.80	7189.43	31971.00	29.00	29087.00	543132.00
Powell	Park	2673.57	1251.00	1.62	3134.18	6407.00	6.00	6314.00	233928.00
Riverton	Fremont	4796.86	2680.00	2.34	5556.49	10953.00	10.00	10615.00	303264.00
Rock Springs	Sweetwater	6620.20	4022.00	2.78	7572.18	24045.00	23.00	23036.00	253584.00
Sheridan	Sheridan	1893.98	2646.00	8.98	6039.71	17916.00	17.00	17444.00	308232.00
sum		33071.38	34064.00	62.80	62652.79	227191.00	210.00	213862.00	3773304.00
average		3006.49	3096.73	5.71	5695.71	20653.73	19.09	19442.00	343027.64
Q1		1829.47	1251.00	1.62	2712.64	6423.00	6.00	6314.00	218376.00
Q3		3894.31	4052.00	8.98	7572.18	31971.00	29.00	29087.00	317736.00
IQR		2064.84	2801.00	7.36	4859.54	25548.00	23.00	22773.00	99360.00
upper fence		6991.58	8253.50	20.02	14861.49	70293.00	63.50	63246.50	466776.00
Lower fence		-1267.80	-2950.50	-9.42	-4576.67	-31899.00	-28.50	-27845.50	69336.00
Outliers				Cheyenne					Cheyenne and Gillette

Outliers


Sheridan
Rock Springs
Riverton
Powell
Gillette
Evanston
Douglas
Cody
Cheyenne
Casper
Buffalo
City
Population Density