11/8/2018

New Store Location for Pawdacity

Predictive Analytics for Business Project

Udacity Nanodegree

Tool: Alteryx

Katerina Bosko, PhD
WWW.CROSS-VALIDATED.COM

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.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

In the second step, you will take the cleaned dataset to train a linear regression model in order to predict sales.

Here are the criteria given to you in choosing the right city:

1. The new store should be located in a new city. That means there should be no existing stores in the new city.
2. The total sales for the entire competition in the new city should be less than $500,000
3. The new city where you want to build your new store must have a population over 4,000 people (based upon the 2014 US Census estimate).
4. The predicted yearly sales must be over $200,000.
5. The city chosen has the highest predicted sales from the predicted set.

Part 1: Data Cleanup

Step 1: Business and Data Understanding

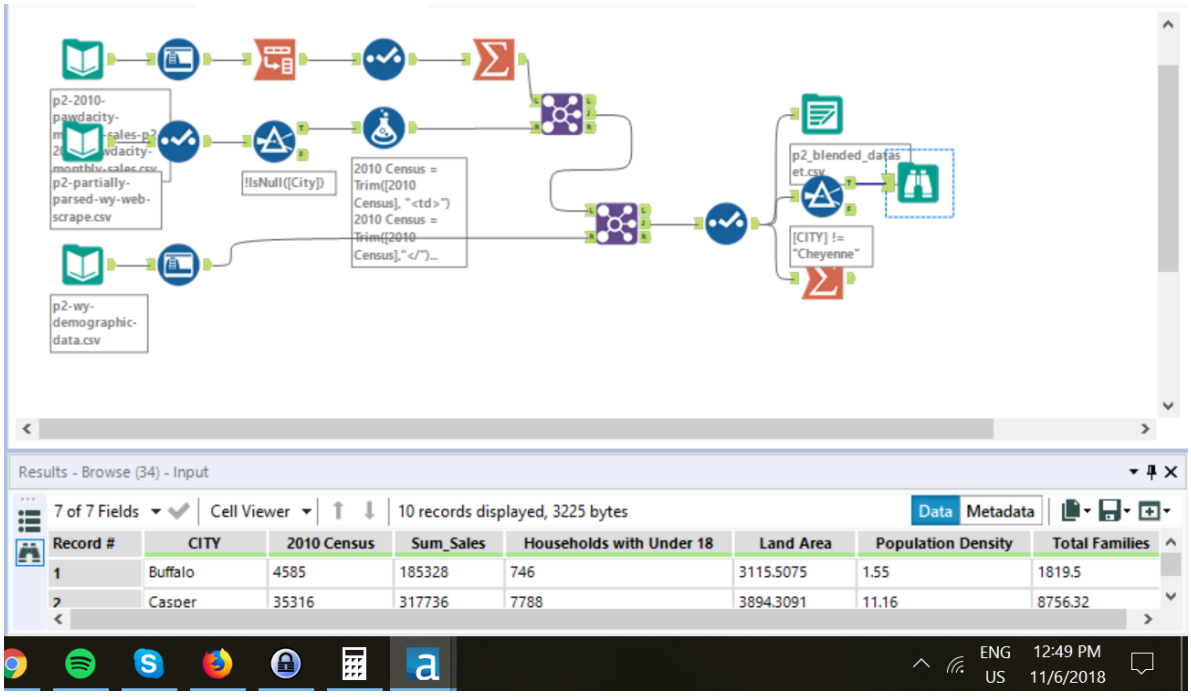
Key Decisions:

1. Pawdacity wants to open a new pet store. The location should be chosen after a thorough analysis of the surrounding area based on the best predicted yearly sales.

2. To inform this decision, we need the Pawdacity stores' yearly sales at the city level. We also need information for other predictor variables such as sale volumes of Pawdacity's competitors in each city, the land area, population density, total families and households with kids under 18 (as families are more likely to have pets and hence to become Pawdacity's customers).

Step 2: Building the Training Set

The Alteryx pipeline for building the training set is provided below:



Column	Sum	Average
Census Population	*213,862*	*19442*
Total Pawdacity Sales	*3,773,304*	*343027.64*
Households with Under 18	*34,064*	*3096.73*

Land Area	*33,071*	*3006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5695.71*

Step 3: Dealing with Outliers

The analysis of outliers using the IQR method in Excel (see picture below) showed that there are three cities with 6 variables that have values beyond interquartile range – Cheyenne, Gillette and Rock Spring. Among these three, I decided to **remove Cheyenne** data from the dataset, because four of its values substantially deviate from the averages (2010 Census, Sum_Sales, Population Density, Total Families) and could bias the results of the regression model to be done in the next step. In general, Cheyenne seems to be a big city with the sales dynamics different from the rest of the cities analyzed in this dataset, which are relatively small. As a result, imputing values for this city doesn't make sense. At the same time, the two other cities with outliers – Gillette and Rock Spring – seem to have the typical "small city" profile and hence, their outliers could be due to chance. As a result, I decided to keep them.

CITY	2010 Cens	Sum_Sale	Household	Land Area	Populatior	Total Fam	2010 Cens	Sum_Sale	Household	Land Area	Populatior	Total Families
Buffalo	4585	185328	746	3115.508	1.55	1819.5	0	0	0	0	0	0
Casper	35316	317736	7788	3894.309	11.16	8756.32	0	0	0	0	0	0
Cheyenne	59466	917892	7158	1500.178	20.34	14612.64	1	1	0	0	1	1
Cody	9520	218376	1403	2998.957	1.82	3515.62	0	0	0	0	0	0
Douglas	6120	208008	832	1829.465	1.46	1744.08	0	0	0	0	0	0
Evanston	12359	283824	1486	999.4971	4.95	2712.64	0	0	0	0	0	0
Gillette	29087	543132	4052	2748.853	5.8	7189.43	0	1	0	0	0	0
Powell	6314	233928	1251	2673.575	1.62	3134.18	0	0	0	0	0	0
Riverton	10615	303264	2680	4796.86	2.34	5556.49	0	0	0	0	0	0
Rock Sprin	23036	253584	4022	6620.202	2.78	7572.18	0	0	0	1	0	0
Sheridan	17444	308232	2646	1893.977	8.98	6039.71	0	0	0	0	0	0

	2010 Cens	Sum_Sale	Household	Land Area	Populatior	Total Families
Q1	7917	226152	1327	1861.721	1.72	2923.41
Q3	26061.5	312984	4037	3504.908	7.39	7380.805
Interquart	18144.5	86832	2710	1643.187	5.67	4457.395
Upper sca	53278.25	443232	8102	5969.689	15.895	14066.9
lower scal	-19299.8	95904	-2738	-603.06	-6.785	-3762.68

Picture – Results of the outlier's analysis in Excel using IQR method

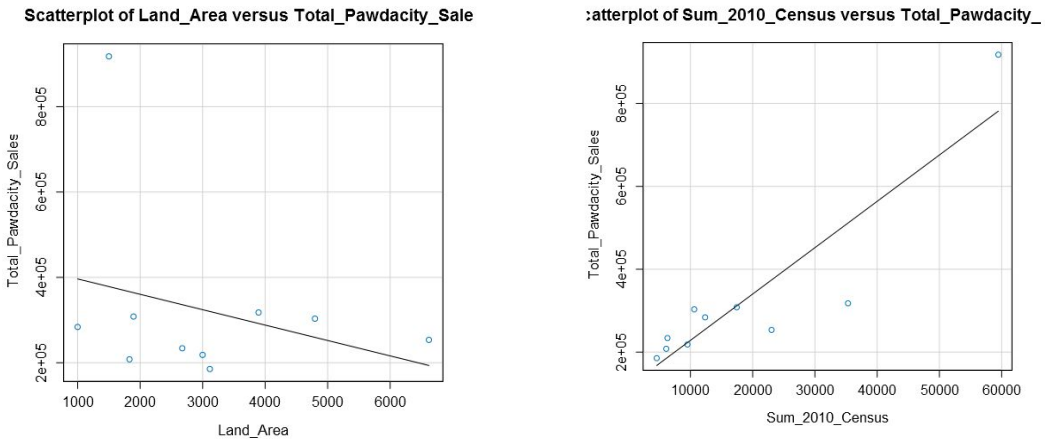
Part 2: Recommend a City

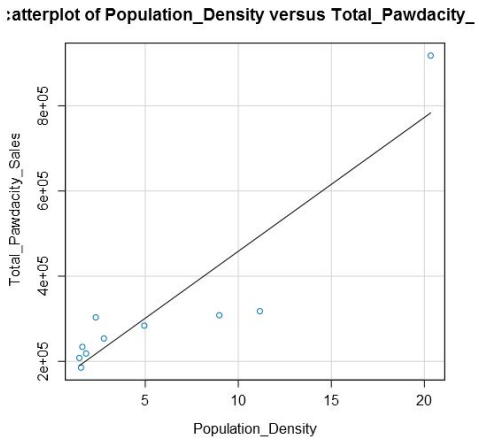
Step 1: Linear Regression

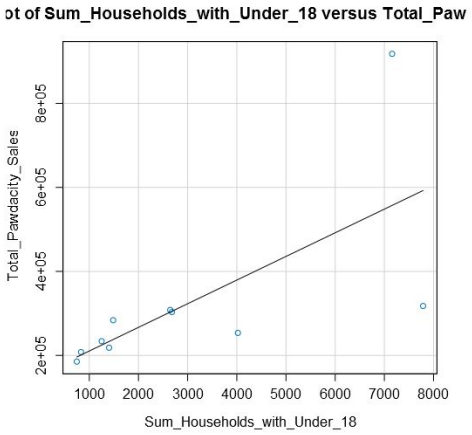
Provide an explanation of the key decisions that need to be made. (250 word limit)

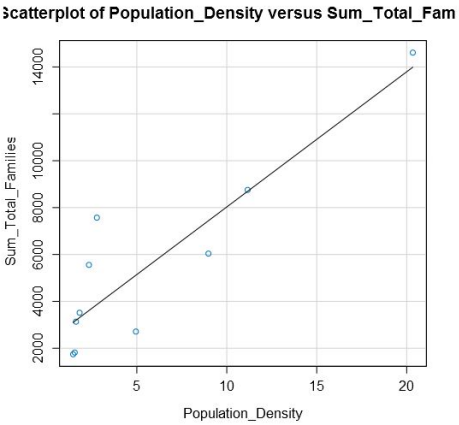
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

I first plotted each predictor variable against my target variable:









I can conclude all predictor variables are good potential predictor variables because they show a linear relationship between sales.

I checked for correlations between my predictor variables to see if there is any possibility of multicollinearity in my dataset. Below is a table that shows the correlations between the different predictor variables:

FieldName	Total Pawdacity Sales	Sum_2010 Census	Land Area	Sum_Households with Under 18	Population Density	Sum_Total Families
Total Pawdacity Sales	1.0000					
Sum_2010 Census	0.8988	1.0000				
Land Area	-0.2871	-0.0525	1.0000			
Sum_Households with Under 18	0.6747	0.9116	0.1894	1.0000		
Population Density	0.9062	0.9444	-0.3174	0.8220	1.0000	
Sum_Total Families	0.8747	0.9692	0.1073	0.9057	0.8917	1.0000

We can see that HHU18, Census, Families, and PDensity (Population Density) have strong correlations which each other. Land area however, is not as highly correlated. So I started by using land area as one predictor and then tested the four variables that are correlated.

I've found out that using land area and total families as the predictor variables produced the best model.

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. You must talk about the p-values and R-squared values that your model produced.

Basic Summary

Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Sum_Total.Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121300	-4453	8418	40490	75200

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197330.41	56449.000	3.496	0.01005	*
Land.Area	-48.42	14.184	-3.414	0.01123	*
Sum_Total.Families	49.14	6.055	8.115	8e-05	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72030 on 7 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866
F-statistic: 36.2 on 2 and 7 DF, p-value: 0.0002035

The p-values for land area and total families are both below 0.05 and the Multiple R-squared value is at .91 which is close to 1. This is model is a decent model.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Y = 197,330 – 48.42 * [Land Area] + 49.14 * [Total Families]

Step 3: Analysis

Use your model results to provide a recommendation. (500 word limit) At the minimum, answer these questions:

1. What kind of data cleaning and aggregation steps did you do?

I started with the Web Scraped Data from the Wyoming Wikipedia page, and used text to columns and select tools and the Data Cleansing to parse out the City, County, 2010 Census, and 2014 Estimate and remove all of the extra punctuation.

For the demographic data, I used the Auto-field tool to combine all of the numbers labeled as String fields.

Before each join, I summarized the amounts by city to ensure that there were no duplicate city names within the data.

For Pawdacity sales file, I transposed the data to get City, Month, and Amount, and then summarized by City to get the total amount for each city.

From there, I created my data set used to train my regression model.

Once the model was created, I applied the model to the cities that were not already in the Pawdacity Sales file by taking the left output from the join on the Pawdacity sales file.

I took the competitor data with an autofield tool and joined it, with a formula off of the left join to create a 0 in the Competitor Amount so I could union the cities that have no competitor back into the overall dataset. I don't want to exclude cities where no competitors are present.

I then applied the filters laid out in the project plan to come up with my list of possible cities, and sorted on the expected revenue to bring the best choice to the top.

2. What were the sales prediction steps did you do?

I filtered my cities according to the given the criteria in the project and calculated revenue off the population density information using my linear model.

3. Which city would you recommend and why did you recommend this city?

I would recommend the city of Laramie with a predicted sales of $305,014.