Recommending Store Location Part 2

Which city would serve as the best location for our new pet store?

Contents

Reporting for Pawdacity:

Ebrahim Tayara
03/28/2018

Overview

Our leading pet store chain in Wyoming has 13 stores throughout the state. This year, they would like to expand and open a 14th store.

Problem

My manager has asked me to perform an analysis to recommend the city for our newest store, based on predicted yearly sales.

Approach

Criteria for choosing the right city:

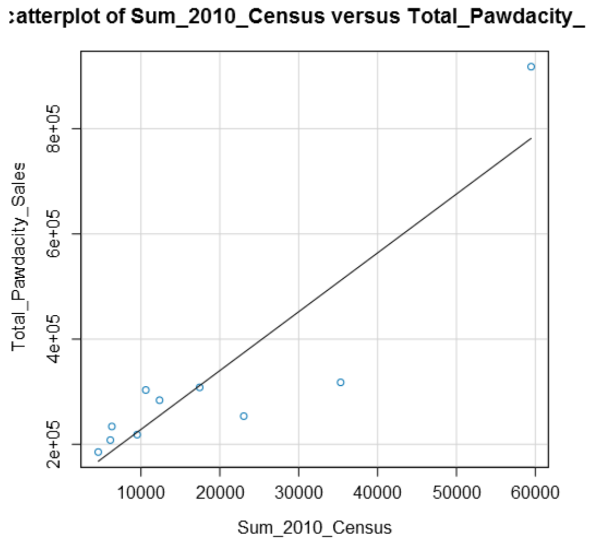
1) The new store should be located in a new city. That means there should be no existing stores in the new city.
2) The total sales for the entire competition in the new city should be less than $500,000.
3) The new city where you want to build your new store must have a population over 4,000 people (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.

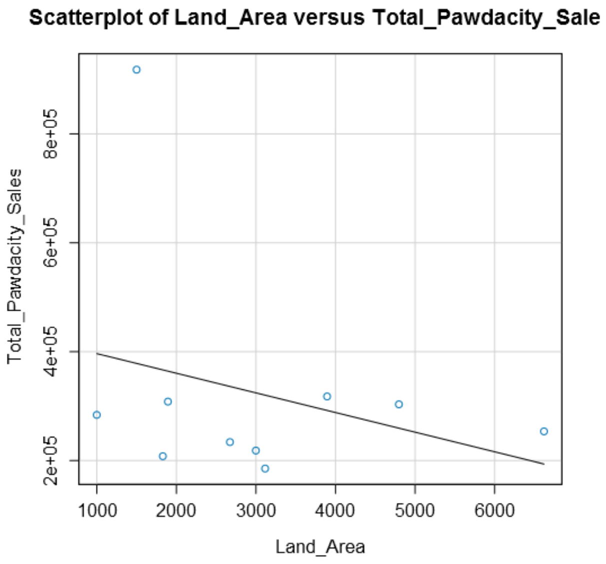
Data Understanding

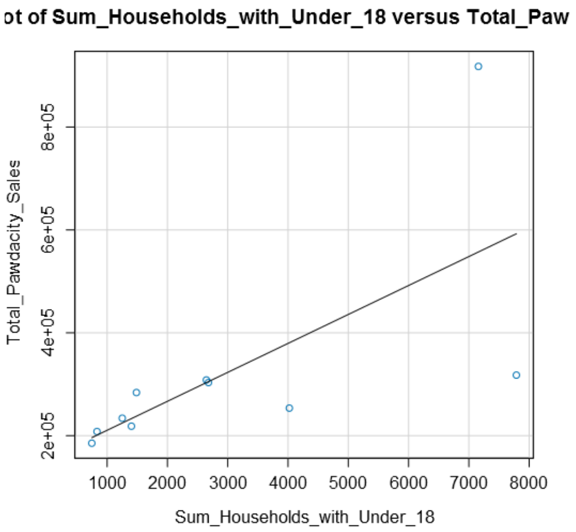
We're using our cleaned-up data from part 1. We found out that Gillette was our biggest outlier, so we removed it from the dataset.

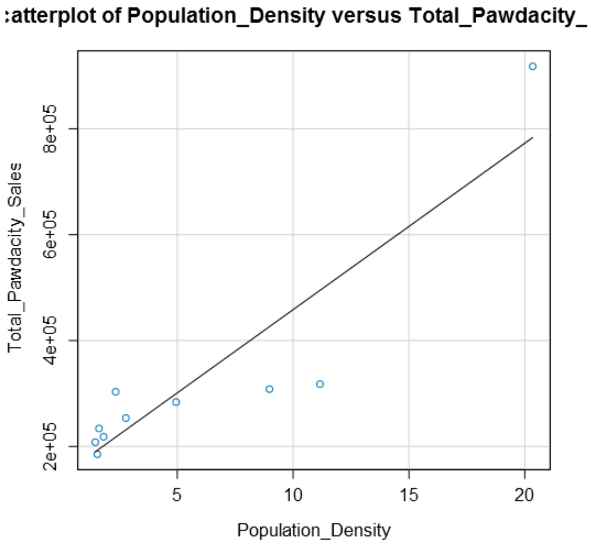
Linear Regression Model

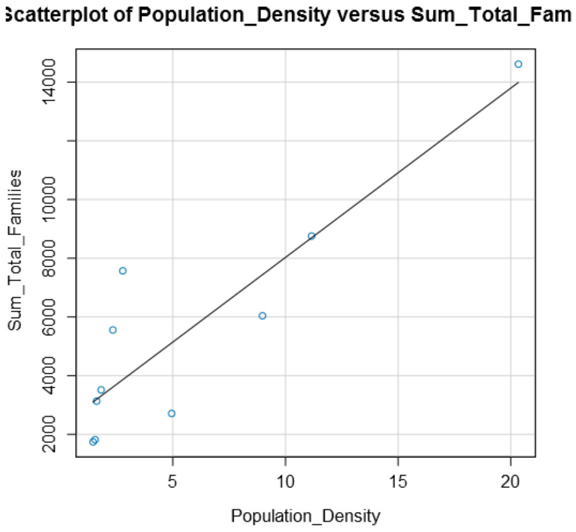
I first plotted each predictor variable against my target variable:





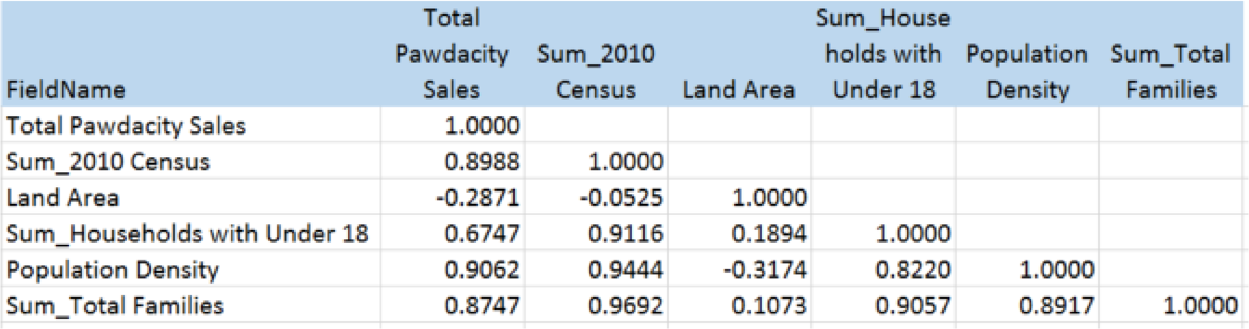






I can conclude all predictor variables are good potential predictor variables because they show a linear relationship between sales.

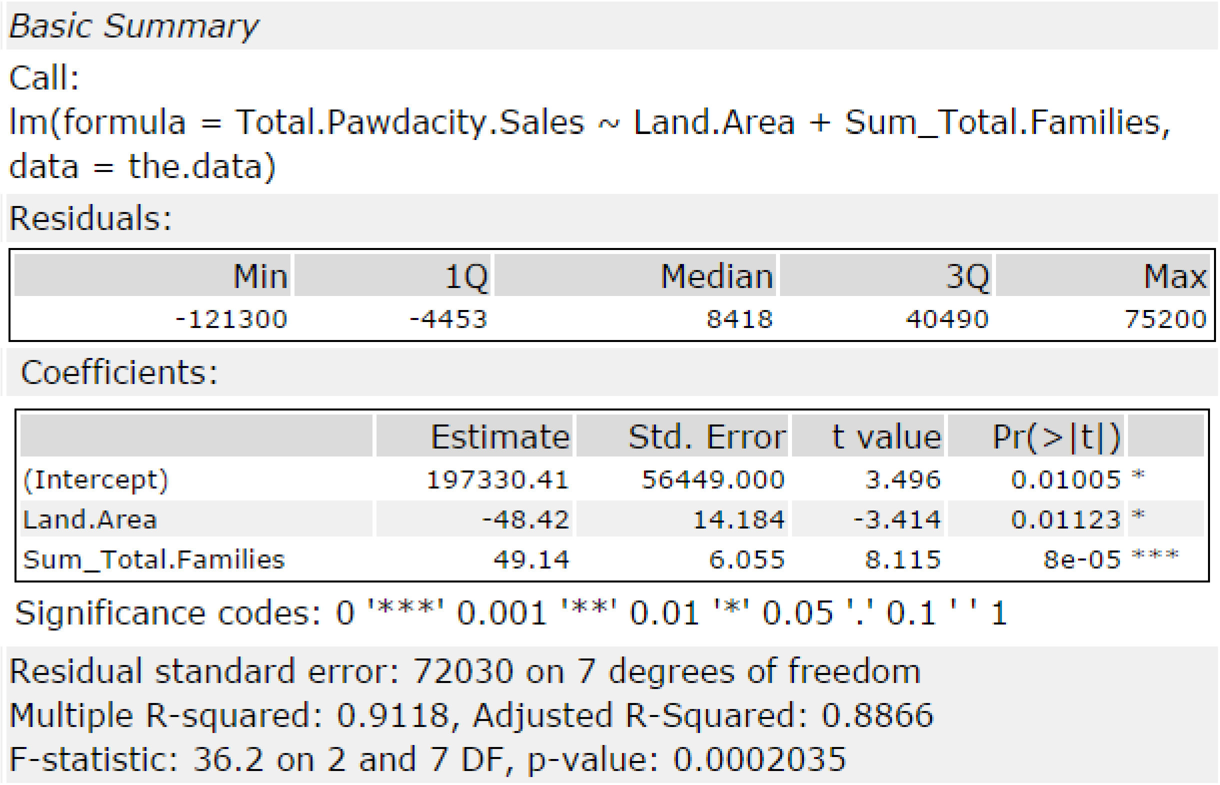
I checked for correlations between my predictor variables to see if there is any possibility of multicollinearity in my dataset. Below is a table that shows the correlations between the different predictor variables:



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.



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

Validation

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.

Please see below for my linear regression equation.

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

Analysis

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 the extra punctuation.

For the demographic data, I used the Auto-field tool to combine all 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 the 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.

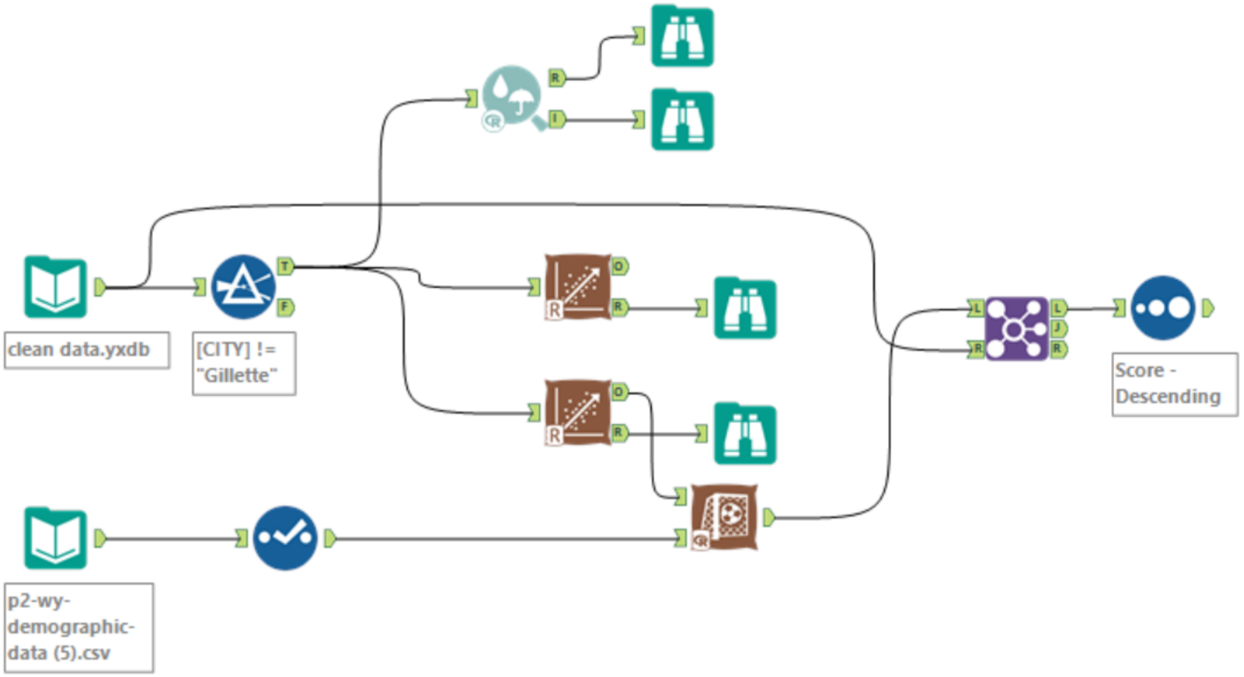
Once the model was created, I applied the model to the cities that were not already in the Pawdacity Sales file by taking the left output from the join on the Pawdacity sales file.

I took the competitor data with an Auto-field tool and joined it, with a formula off 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 filters to come up with my list of possible cities and sorted on the expected revenue to bring the best choice to the top.

Finally, to predict sales, I filtered my cities according to the given criteria and calculated revenue off the population density information using my linear model.

Workflow



Decision

I recommend we set up our new store in the city of Laramie due to predicted sales of ~ $305,014.

Sources:

Compiled from data obtained from Udacity.

Update History:

1. Created – 28-March-2018.