Selecting the Location for a Pet Store

By Jannet Popescu

Part 1 - Project Overview

This project is the first part of a *two-part series*. In the first part, I will be working in data cleanup, formatting the data, and dealing with outliers. For the second part, I will use the cleaned up dataset to create another linear regression model. Then, I will choose which variable(s) are the most important for the model.

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. My job is to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

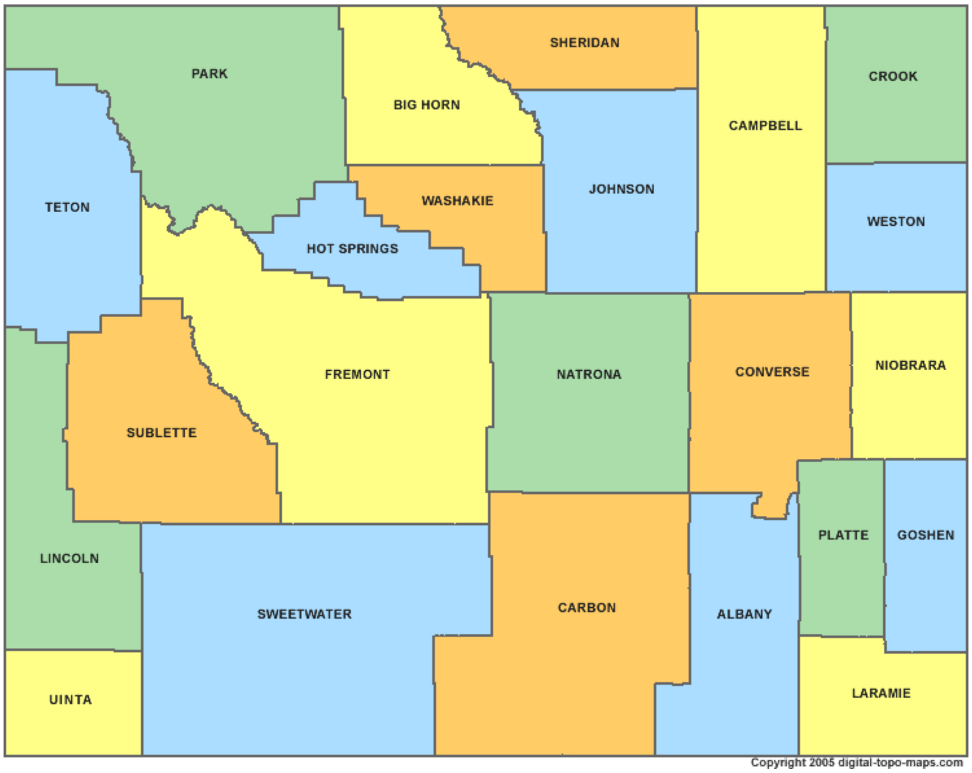
My first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Details

The following information has been provided:

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.

Map of Wyoming Counties



Part 1 - Project Submission and Findings

Step 1: Business and Data Understanding

What decisions needs to be made?

Pawdacity would like to expand and open a 14th store in Wyoming. We need to decide the best location to build the new store, and if the projected target sales will exceed $200,000 in the first year of operations.

What data is needed to inform those decisions?

What data is available?

We have Pawdacity's 2010 monthly sales and stores locations. Wyoming demographic data including census data (total households with people under 18, population density, total families), city, county and land area. Additionally, we have the competitors market data in Wyoming.

What data is needed?

It would be useful to have data on projected population growth in the city, the number of pets owned per household, and type of pet (cat, dog, etc.). We'll also require additional data about our competitors such as marketing budget, total store expenditures per city, and current local promotions. Additionally, having more information about location of recreational facilities such as parks, shared pet spaces, etc. would be helpful to narrow down our location search but is not necessary for our analysis.

Step 2: Building the Training Set

To properly build the model, and select predictor variables, I've create a dataset with the following columns:

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

This dataset will be my training set to help you build a regression model in order to predict sales in Project 2.2. Every row should have sales data because we're trying to predict sales.

Notes

I've consolidated the data at the city level and *not at the store level.* We only have data at the city wide level so any analysis at the store level will not be sufficient to complete this analysis. I've focused on cleaning up and blending the data together in this step.

Column	Sum	Average
Census Population	213,862	19,442.00
Total Pawdacity Sales	3,773,304	343,027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.45
Population Density	63	5.73
Total Families	62,653	5,695.73

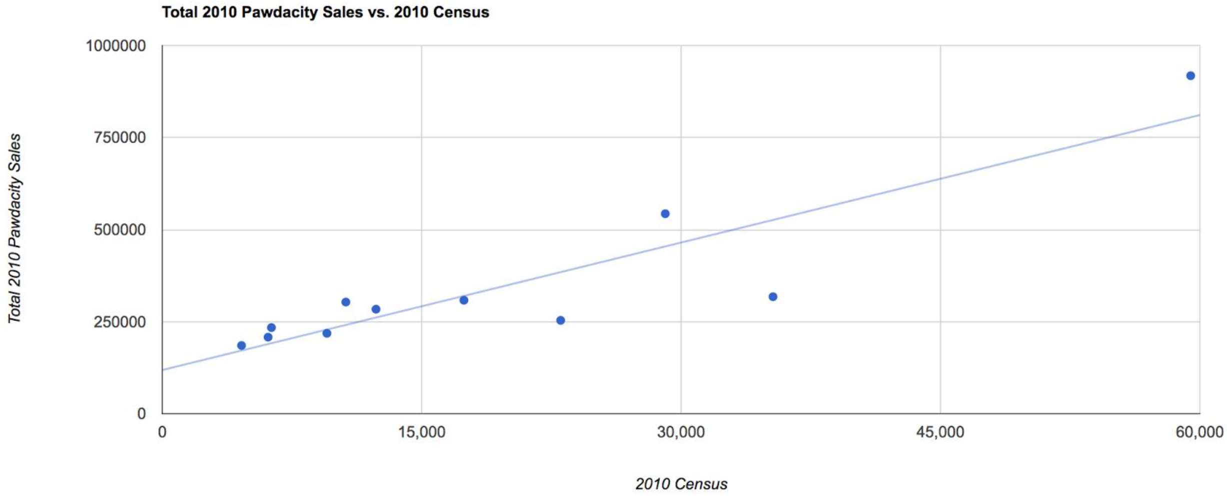
Step 3: Dealing with Outliers

Are there any outliers in the training set? Because the dataset is a small data set (11 cities), I've chosen to remove or impute one outlier. Below is my explanation analysis.

Land Area does not have outliers based on the interquartile analysis of the data set. Similarly, **Households with Under 18** does not have outliers, which seems to follow our reasoning that households with a higher number of members under 18 will have a higher likelihood to purchase items from Pawdacity.

Population Density presents one outlier for the largest city in Wyoming - *Cheyenne*, which has a population density of 20.34. However, there is a relationship between population density and sales based on a scatter plot analysis, so this outlier is not skewing our data. In addition, this outlier will be useful to build more robust predictive analysis for future modeling analysis for store expansion in big cities. We are keeping Cheyenne as part of our analysis. Similarly, **Total Families** also presents an outlier -- *Casper*. However, the city does not skew our findings in terms of sales, so we are keeping Casper as part of the analysis.

Census Population presents an outlier we can exclude -- *Gillette*. Using the interquartile range analysis, the sum_sales for Gillette appears as our clearest outlier. Although, our census population data set does not skew our analysis, we have decided to leave Gillette out of the model analysis due to this data point being an outlier that goes against our logic towards achieving a target sales goal of over $200,000, which is likely to be achieved if the 14th store is established in a larger city.



Part 2 - Project Overview

This project is a continuation of Project 1 regarding trying to find the best city to expand for Pawdacity's newest pet store.

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. My job is to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

In the first part, I've already cleaned up the dataset and dealt with outliers. In Part 2 of this project, I will take the dataset that I've cleaned up and use it to train a linear regression model in order to predict sales.

Details

Here are the criterias given 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 the store will be built 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 2 - Project Submission and Findings

Step 1: Linear Regression Model

I'll be building a linear regression model to help me predict total Pawdacity sales for the 14th store location. I will be analyzing the dataset created in Project 2.1 and look at the distribution of the data. I have 10 rows of data before modeling the dataset. I've removed **Gillette** from the dataset since it's an outlier.

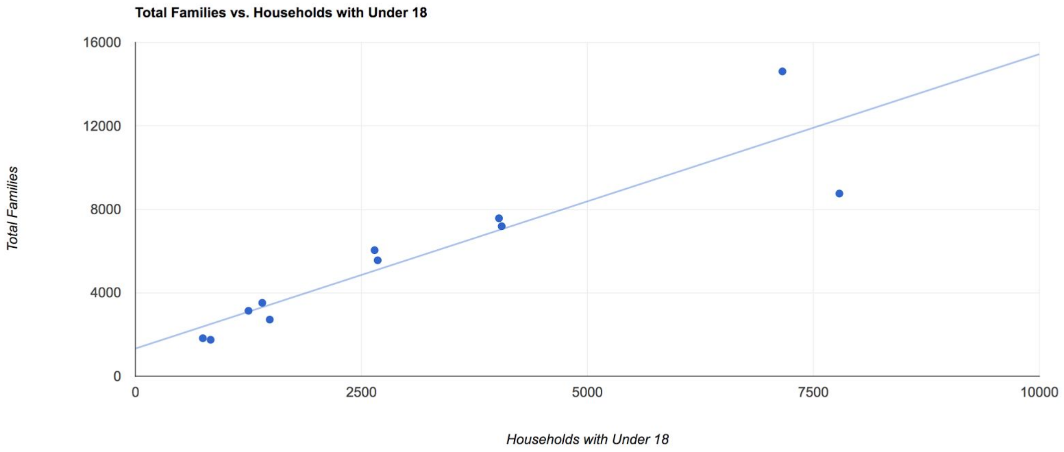
In developing the linear regression model, the first thing to verify is the linear relationship between the target variable and each predictor variable. Next, I'll be looking at the correlation coefficients of each predictor variable to see if there's any possibility of multicollinearity.

Pearson Correlation Analysis

Full Correlation Matrix

	2010_Census	Land Area	Households_with_ Under_18	Population Density	Total Families
2010_Census	1.000000	0.011028	0.978353	0.898532	0.978489
Land Area	0.011028	1.000000	0.189376	-0.317419	0.161786
Households_with_ Under_18	0.978353	0.189376	1.000000	0.821986	0.996996
Population Density	0.898532	-0.317419	0.821986	1.000000	0.835744
Total Families	0.978489	0.161786	0.996996	0.835744	1.000000

The correlation analysis above shows that our data (2010_Census, Households_with_Under_18, Total Families, and Population Density) have strong correlations with each other. There is multicollinearity involved, so our next step is to find the best regression model by simple looking at each individual predictor variable.



It appears that Households_with_Under_18 and Total Families are highly correlated to one another. For this reason, one of these predictor variables will be removed, and post removal the new linear regression model will look as next:

Pawdacity Linear Regression Model

Basic Summary

Call: lm(formula = Sum_Sales ~2010_Census + Land_Area + Households_with_Under_18 + Population_Density, data = the data)

Residuals

Min	1Q	Median	3Q	Max
-132000	-50160	924	52130	108500

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30636.55	109356.67	0.2802	0.79058
2010_Census	10.81	15.82	0.6833	0.52481
Land_Area	62.14	44.58	1.3937	0.22218
Households_with_Under_18	-138.46	122.52	-1.1301	0.30972
Population Density	52315.29	18217.19	2.8718	0.03492

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

Residual standard error:	98427 on 5 degrees of freedom
Multiple R-squared:	0.8824
Adjusted R-squared:	0.7883
F-statistic:	9.38 on 4 and 5 DF
p-value:	0.0152

Based on Pawdacity's initial linear regression model, it appears the Population Density is the only significant predictor variable with a p-value less than 0.05. For this reason, all other predictor variables will be omitted, and we will be focusing on re-running a regression model only using Population Density as our predictor variable. See the new regression model below:

Pawdacity Linear Regression Model - *Population Density*

Basic Summary

Call: lm(formula = Sum_Sales ~ Population_Density, data = the data)

Residuals

Min	1Q	Median	3Q	Max
-177000	-13510	17830	34990	134600

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	143800	42371	3.394	0.00945
Population Density	31442	5188	6.061	3e-04

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

Residual standard error:	95963 on 8 degrees of freedom
Multiple R-squared:	0.8212
Adjusted R-squared:	0.7988
F-statistic:	36.73 on 1 and 8 DF
p-value:	0.0003023

Our new regression analysis with an R-squared of 0.8212 (greater than our threshold of 0.7), represents a strong model. For this reason, we will use Population Density to build our regression equation and predict sales for future Pawdacity locations in Wyoming.

Regression Equation

Sum_Sales = 143,800 + 31,442 * Population_Density

Step 2: Analysis

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

In order to make a recommendation, we first need to consider the following information:

- The new store should be located in a new city where Pawdacity does not have an existing store.
- The total sales for the entire competition in the new city should be less than $500,000.
- The new city where Pawdacity wants to build a store much have a population of over 4,000 people (based upon the 2014 US Census estimate data).
- The predicted yearly sales must be over $200,000.
- The city chosen must have the highest predicted sales from the predicted set.

Based on these conditions and after calculating in our dataset the total predicted sales in cities in Wyoming. We have narrowed down our search for the next city where Pawdacity should open its 14th store. The city is **Laramie** with a predicted annual sales of **$306,983.98**.