Predictive Analytics for Business

Project#3-2 Create an Analytical Dataset

Name: Marwan Saeed Alsharabbi

Date: 27-12-2019

2019

INTRODUCTION

.

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.

In the first part, you've already cleaned up the dataset and dealt with outliers.

In this project, you will take this dataset that you cleaned up and use this dataset to train a linear regression model in order to predict sales

Here are the criteria's 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.

Steps to Success

Step 1: Build a Linear Regression Model

Analyze the dataset you created in Project 2.1 and look at the distribution of your data. You can create histograms to look at each of your continuous and categorical data to determine the nature of the data you're working with.

Important: You should have **10 rows** of data before you begin modeling the dataset. The correct answers will assume that you removed **Gillette** from the dataset. If you decided to remove a different outlier in P2.1, you will have to adjust your dataset for this project.
Build a linear regression model to help you predict total sales.

Step 2: Perform the Analysis

Use your regression model to calculate predicted sales for all of the cities and use the criteria given to you to make a recommendation.

Project3.2: Practice Project: Recommend a City

Note that this project is a continuation of the Data Cleanup project.

Step 1: Linear Regression

Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

Important: *Make sure you have dealt with outliers and removed one city from your training set. You should have* ***10 rows*** *of data before you begin modeling the dataset.*

Build a linear regression model to help you predict total sales.

At the minimum, answer these questions:

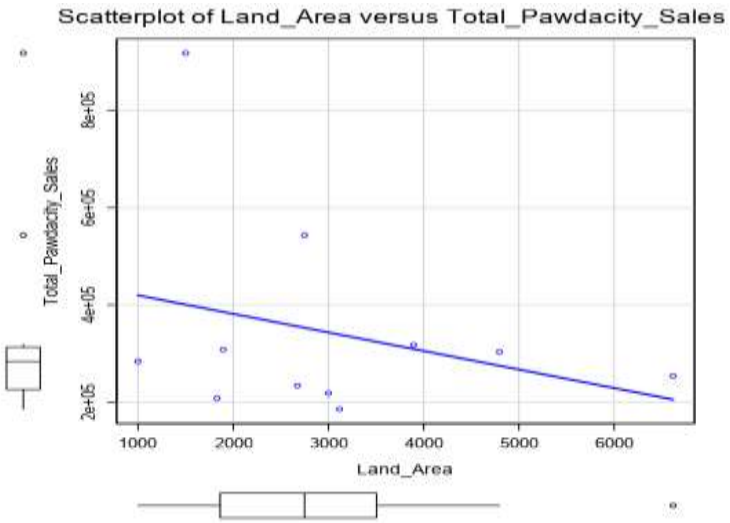
1-How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

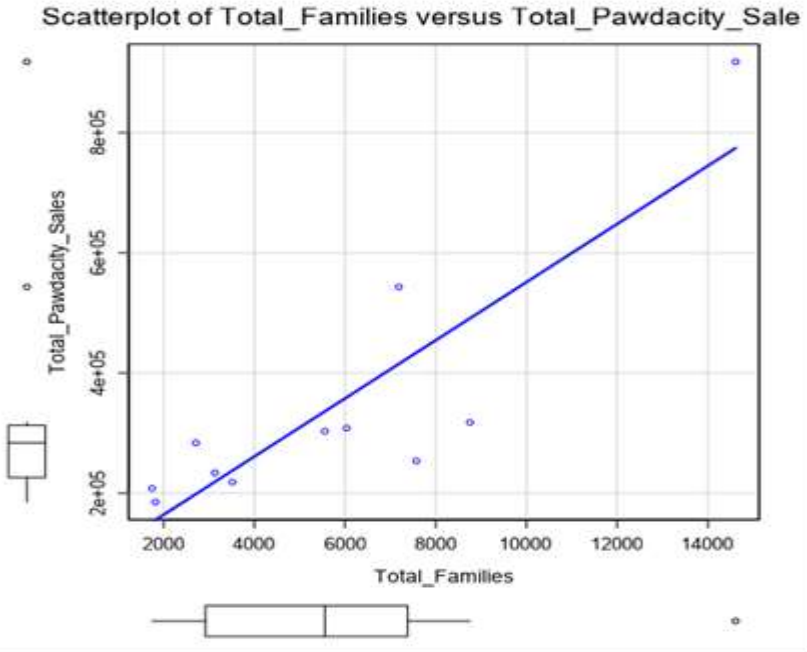
Data Sets use in Project

Data Understanding

- What data is needed?
- What data is available?
- What are the important characteristics of the data?

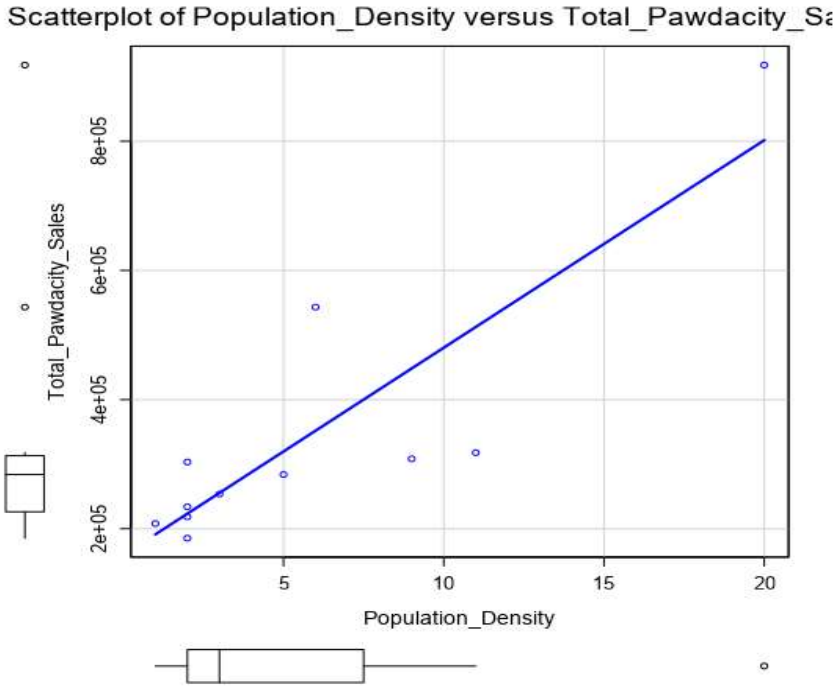
In this project the problem is explained and the data are available and I will choose the necessary data tables in order to clean them by using a program Alteryx
I'm chose three sets data:
p2-2010-pawdacity-monthly-sales.csv - This file contains all of the monthly sales for all Pawdacity stores for 2010.
p2-partially-parsed-wy-web-scrape.csv - This is a partially parsed data file that can be used for population numbers.
p2-wy-demographic-data.csv - This file contains demographic data for each city and county in Wyoming.
In the first part, you will blend and format data and deal with outliers.

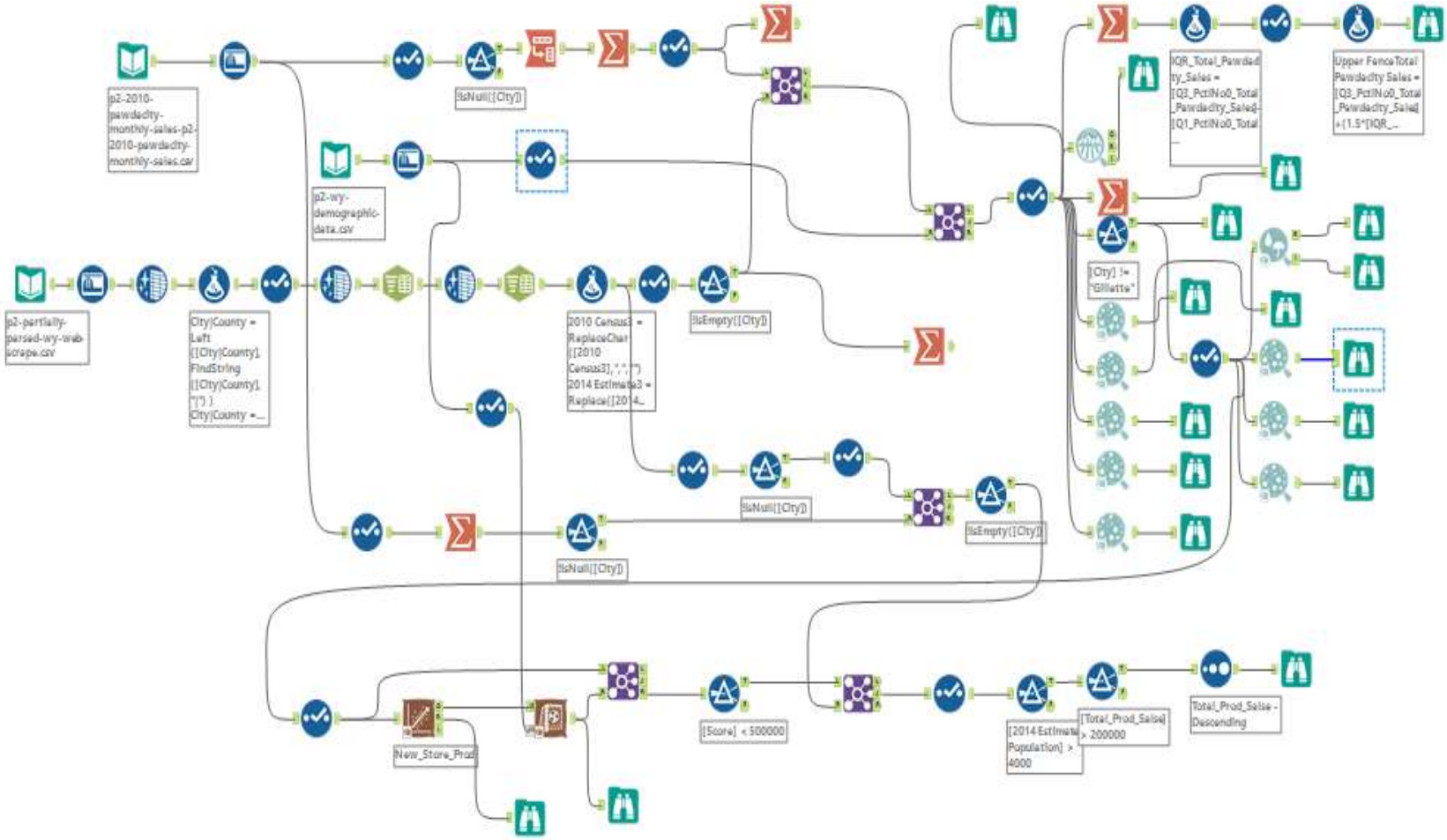




I first plotted each predictor variable against my target variable

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.
The scatter plots give a good representation of the linearity between the target variable and its respective predictor variable.





There are two ways to implement multiple linear regression either Ba import data after the cleaning process or direct linkage so that the project steps from the interconnected beginning to be the end

Report for the linear Model New_Store_Prod

Report

Report for Linear Model New_Store_Prod

Basic Summary

Call:
lm(formula = Total_Pawdacity_Sales ~ Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121260	-4467	8422	40490	75208

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197299.27	56451.744	3.495	0.01006	*
Land_Area	-48.41	14.184	-3.413	0.01124	*
Total_Families	49.13	6.055	8.115	8e-05	***

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

Residual standard error: 72033 on 7 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866
F-statistic: 36.2 on 2 and 7 degrees of freedom (DF), p-value 0.0002035

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)	
Land_Area	60453713643.39	1	11.65	0.01124	*
Total_Families	341664344221.7	1	65.85	8e-05	***
Residuals	36321013347.65	7			

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

1-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. . For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

I'm selected the target variable Total_Pawdacity_Sales and Land_Area and Total_Families as my predictor variables for my linear R-Squared comes from Land_Area and Total_Families (adjusted r-squared = 0.8866) I'm use Land_Area and Total_Families as my predictor variables for my linear 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.

2-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 (Total_Pawdacity_Sales) = 197,229.27 – 48.41 * [Land Area] + 49.13 * [Total Families]

Step 2: Analysis

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer this question:

I started with the Web Scraped Data from the Wyoming data sets from project, 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.

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

With the required criteria, I would recommend **Laramie City**. Laramie City does not currently contain a store, has an 2014 estimated population for 2014 of 32,081 and Total_Families is 4668.93

Total_pred_sales of **$305,004**

Summary of the final possibilities for a new store

Record	City	Land_Area	Total_Families	Total_Prod_Salse	2014 Estimate Population
1	Laramie	2513.745235	4668.93	305004	32081
2	Torrington	1599.818493	2548.5	245064	6736
3	Jackson	1757.6592	2313.08	225855	10449
4	Lander	3346.80934	3876.81	225750	7642
5	Green River	3477.361206	3977.4	224372	12630
6	Worland	1294.105755	1364.32	201681	5366

Help resources :https://knowledge.udacity.com

I wish success to all.

M*arwan Saeed Alsharabbi*