Project 2.2 – Predict location of a new Pet store

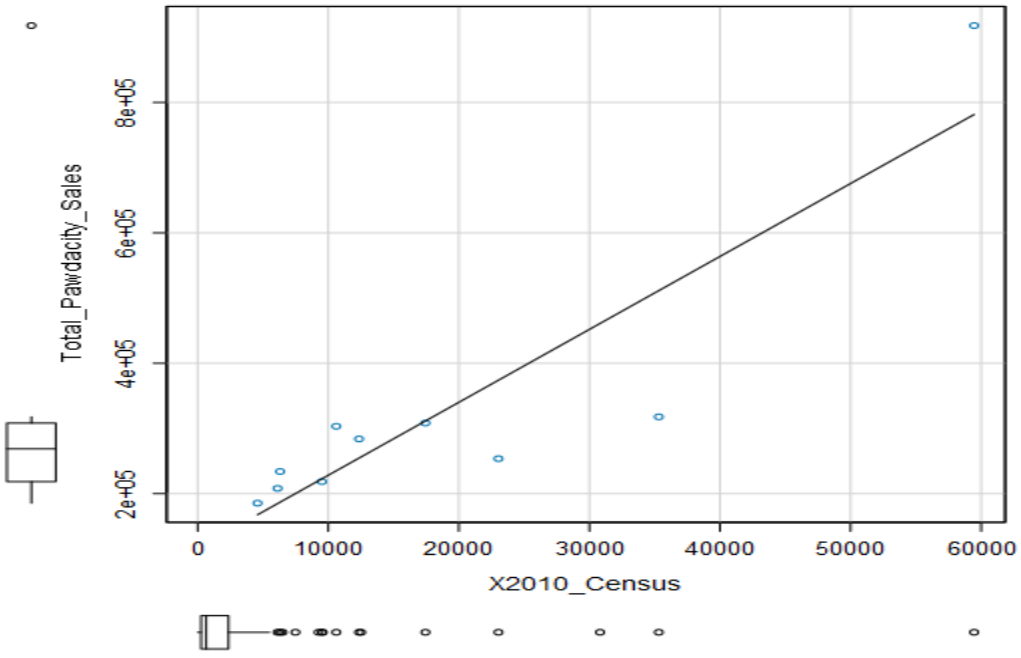
Chandan Mishra

Step1: Linear Regression

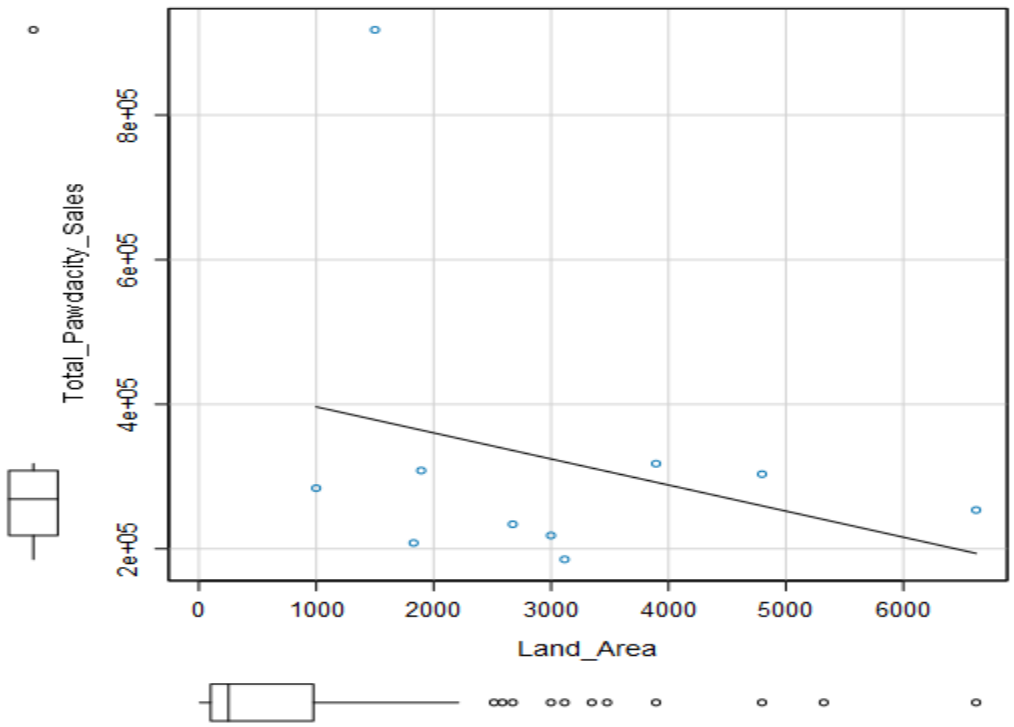
1. **How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.**

Plotting each predictor variable against the target variable (Total pawdacity sales):

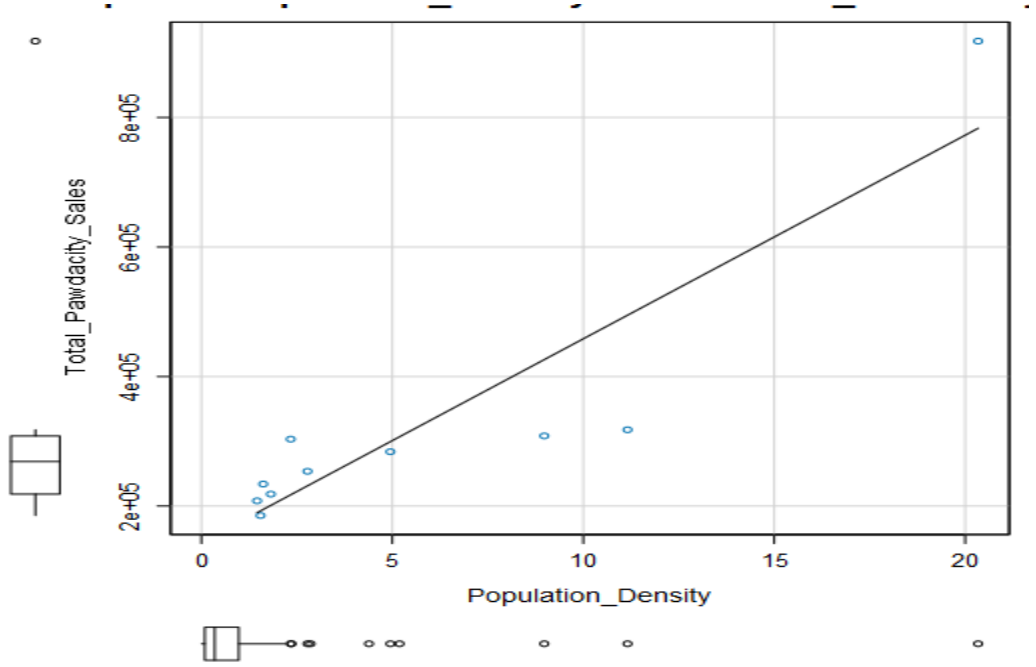
a) Total Pawdacity sales vs 2010 census



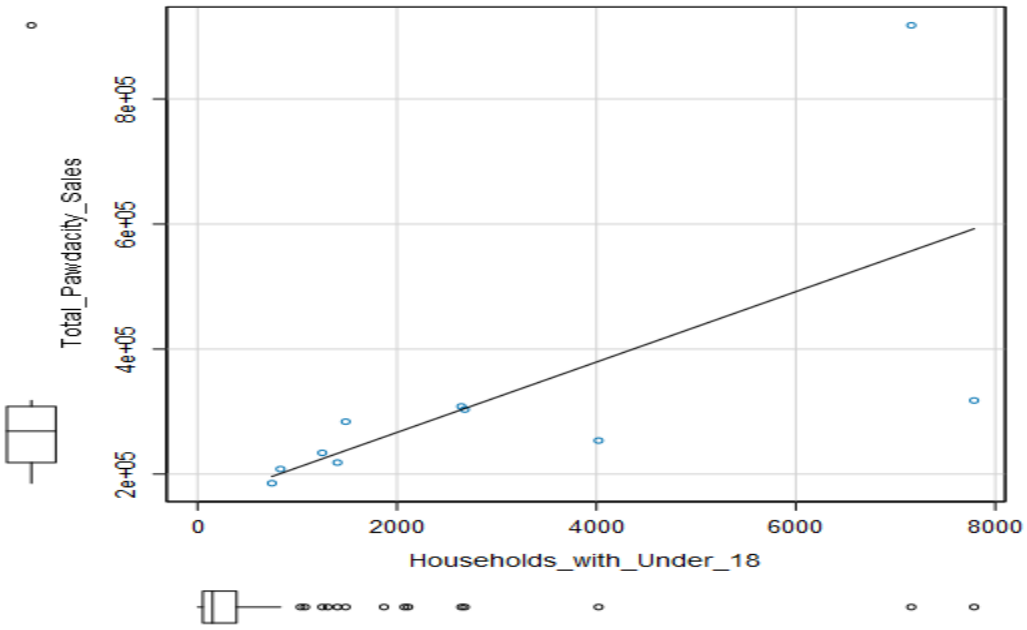
b) Total Pawdacity Sales vs Land Area



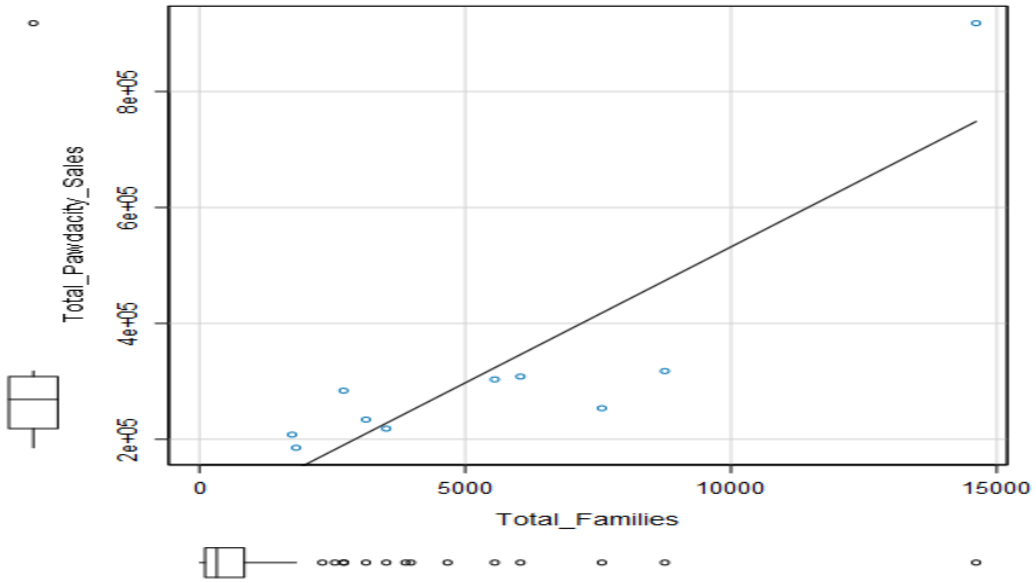
c) Total Pawdacity Sales vs Population density



d) Total Pawdacity Sales vs Household with Under 18



e) Total Pawdacity Sales vs Total Families



All these predictor variables are strong predictors since they share a linear relationship with the target variable. To check correlation between the predictor variables, I ran an association analysis which is shown as below:

Field Name	Total_Pawdacity_Sales	2010_Census	Land_Area	Households_with_Under_18	Population_Density	Total_Families
Total_Pawdacity_Sales	1					
2010_Census	0.9	1				
Land_Area	-0.29	-0.05	1			
Households_with_Under_18	0.67	0.91	0.19	1		
Population_Density	0.91	0.94	-0.32	0.82	1	
Total_Families	0.87	0.97	0.11	0.91	0.89	1

As can be seen the ones marked in red are the least correlated while the ones marked in green or yellow are highly correlated with each other. Land area can be chosen as a predictor variable over here along with the combination of the other four variables related to population or families.

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. . 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.

Testing the combination one by one, I found Land_Area and Total_Families to produce the best model.

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-121261	-4453	8418	40491	75205

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 *
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 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	60473052720.43	1	11.66	0.01123 *
Total_Families	341673845917.83	1	65.85	8e-05 ***
Residuals	36318449406.44	7		

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

p-value for land_Area is 0.01123 and for total_Families is 8e-05 which means both are less than 0.05 and are statistically significant. The Adjusted R square value i.e. 0.88 is also very high which represents the model is a good model.

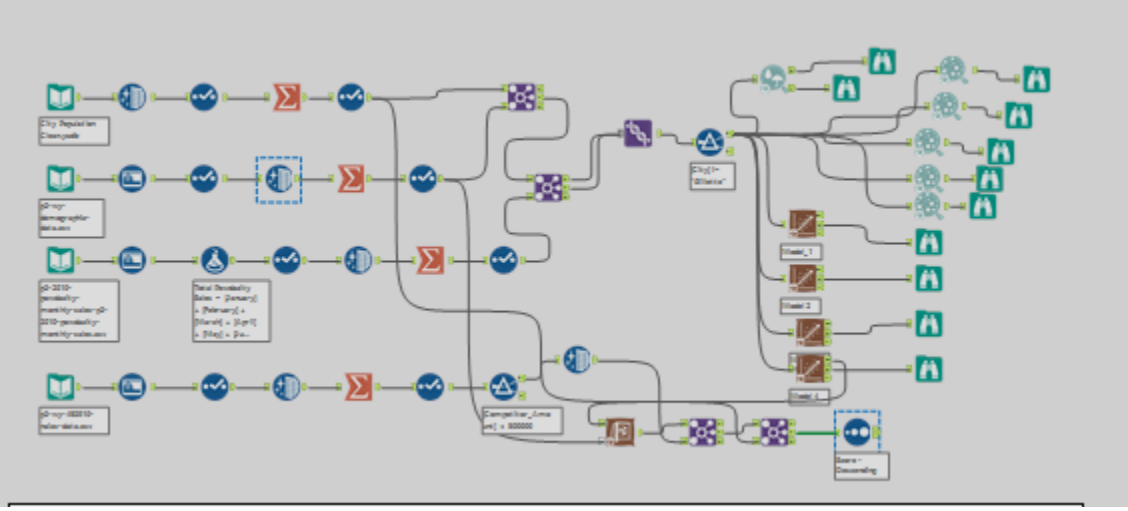
3. **What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)**

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

Step 2: Analysis

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

1. **What kind of data cleaning and aggregation steps were taken?**



The model looks like this.

City Population data, which was already cleaned in part 1 of this project was summarized at the city level.

Demographics data was also cleaned and summarized at the city level.

Both the tables were then joined. Let's name this table 1.

Pawdacity sales data was also summarized at the city level and was joined with Table 1. To not loose any cities, the left and right data after join was unioned and City "Gillette" because of being an outlier was removed. After running association analysis and checking the scatter plots of all the predictor variables, various models were run. Model 4 where the predictor variables were Land Area and Total Families produced the best result. To predict pawdacity sales for the cities, the model was put in a score tool using data from the demographics. Only cities with competitors' sales less than $500,000 were kept in the model. After sales were predicted, it was sorted in a descending sequence to find out about the city with the most predicted sales.

2. **Which city would you recommend and why did you recommend this city?**

Record #	City	Land_Area	Households_with_Under_18	Population_Density	Total_Families	Score	Competitor_Amount	2014_Estimate	2010_Census
1	Casper	3894.3091	7788	11.16	8756.32	438997.172236	210000	40086	35316
2	Laramie	2513.745235	2075	5.19	4668.93	305013.881671	76000	32081	30816

Since Casper city already has a pawdacity store, so we will ignore the first record and select Laramie as the potential new location based on Total predicted sales which is $305,013.88.