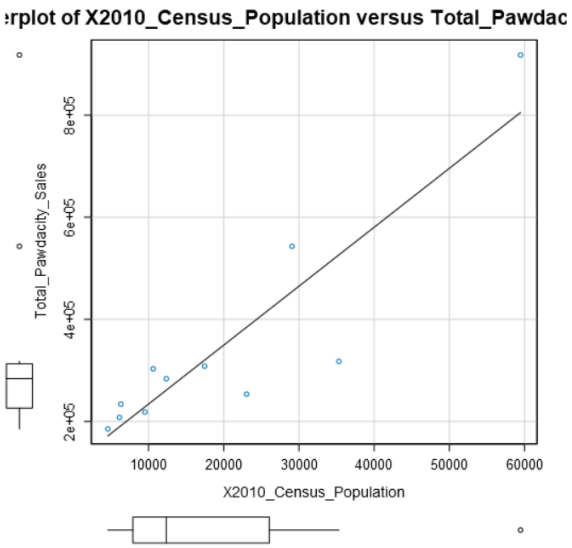
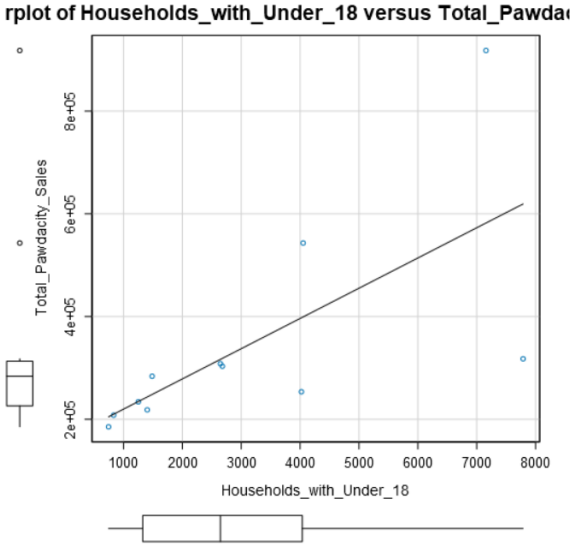
Project 2.2: Recommend a City

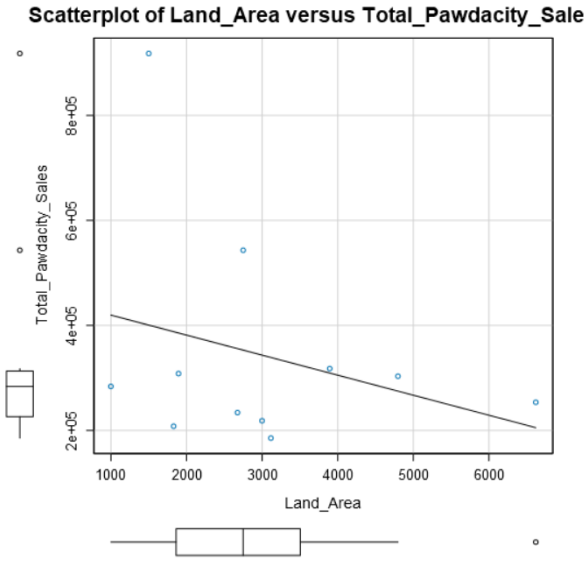
Step 1: Linear Regression

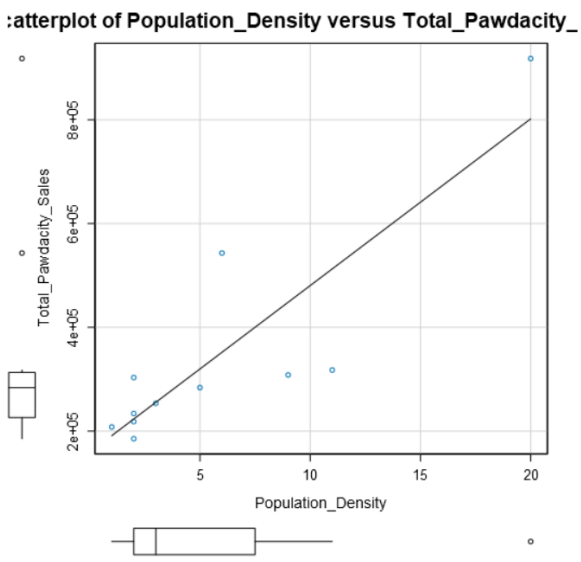
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

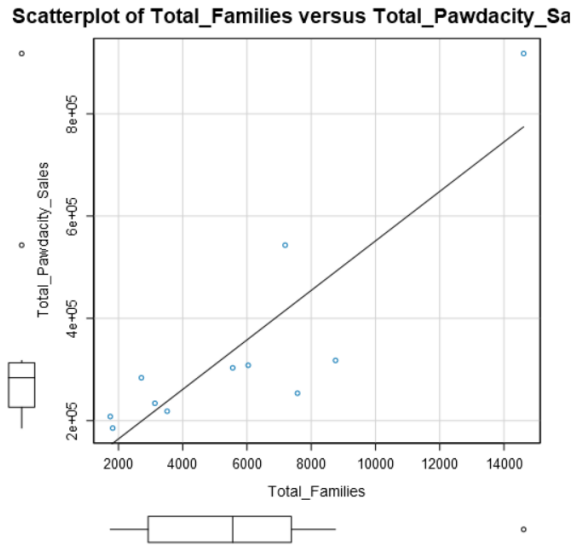
As we can see in scatterplots for all potential predictor variables. They all show linear relationship between sales. Note: All being positive linear relationship except Land Area.









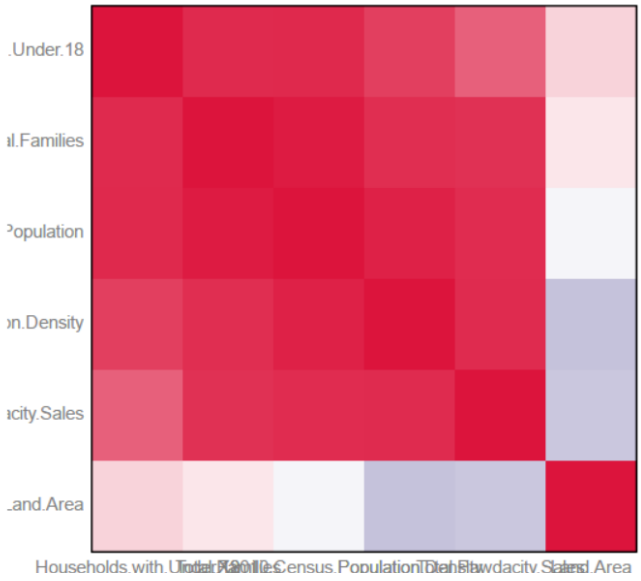


I've used Association Analysis Tool to see the correlation between these variables and Total Sales.

Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

	Association Measure	p-value
Population.Density	0.90185	0.00036008 ***
X2010.Census.Population	0.89875	0.00040617 ***
Total.Families	0.87469	0.00092495 ***
Households.with.Under.18	0.67465	0.03235537 *
Land.Area	-0.28711	0.42121354



Population Density, Census Population, Total Families, and Households with Under 18 are highly correlated to Total Sales as they all have p-value <= 0.05. However, I would guess they are also highly correlated with each other. So I picked Land Area as predictor variable and experimented with Population Density, Census Population, Total Families, and Households with Under 18.

After the experimentation, I've found out that using Land Area and Total Families are the predictor variables produced the best model.

Report for Linear Model Linear_Regression_11

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-121300	-4467	8422	40490	75210

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 DF, p-value: 0.0002035

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.

The p-values for Land Area and Total Families are both below 0.05 and the Multiple R-squared value is at 0.91 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?

Y = 197,299 – 48.41 * [Land Area] + 49.13 * [Total Families]

Step 2: Analysis

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

I would recommend the city of Laramie with a predicted sale of $305,004.

Record #	City	County	Score
1	Laramie	Albany	305003.767733

Here is the validation for the criteria:

1. The new store should be located in a new city. That means there should be no existing stores in the new city.

Record #	City
1	Buffalo
2	Casper
3	Cheyenne
4	Cody
5	Douglas
6	Evanston
7	Gillette
8	Powell
9	Riverton
10	Rock Springs
11	Sheridan

2. The total sales for the entire competition in the new city should be less than $500,000

   The sales volume for the only competitor in Laramie is $76,000

Record #	BUSINESS NAME	PHYSICAL CITY NAME	SALES VOLUME	CASS_LastLine
1	Muddy Paws Pet Salon	Laramie	76000	Laramie, WY 82070-8979

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).

   The population of Laramie in 2014 US Census estimate is 32,081.

Record #	2014 Estimate	City	Country
1	32,081	Laramie	Albany

4. The predicted yearly sales must be over $200,000.

   Predicted sale is $305,004.

5. The city chosen has the highest predicted sales from the predicted set.

Record #	City	County	Score
1	Laramie	Albany	305003.767733
2	Torrington	Goshen	245064.414204
3	Mills	Natrona	239617.688583
4	Evansville	Natrona	229766.551195
5	Bar Nunn	Natrona	228665.132705
6	Jackson	Teton	225855.25272
7	Lander	Fremont	225750.388473
8	Green River	Sweetwater	224372.211136
9	Lyman	Uinta	219661.093607
10	Wright	Campbell	218284.779507
11	Pine Bluffs	Laramie	217780.958355
12	Wheatland	Platte	214240.513717