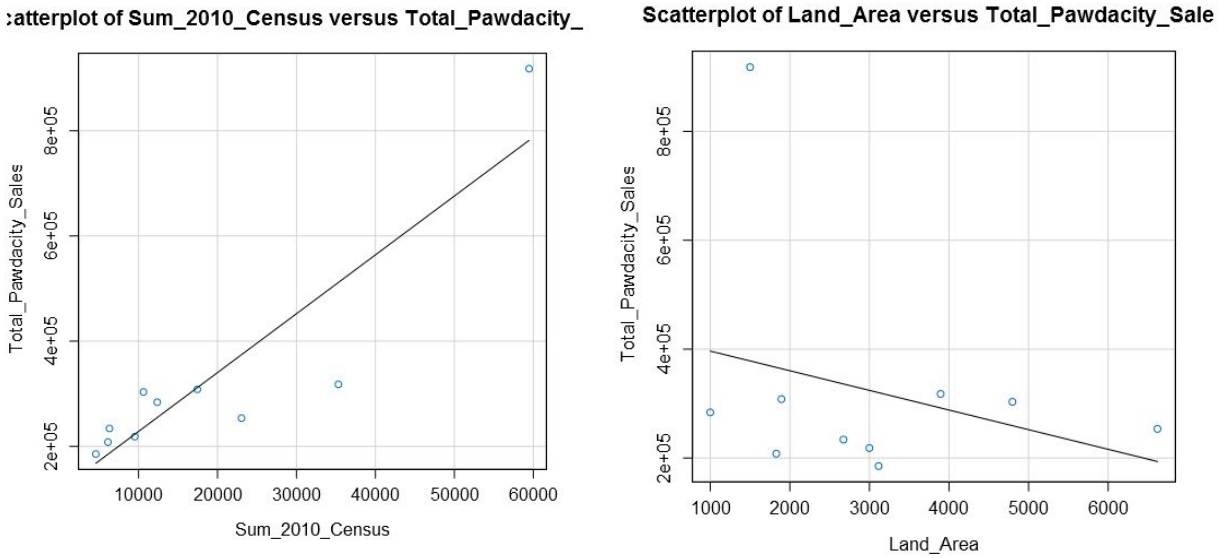
Project 2.2: Recommend a City

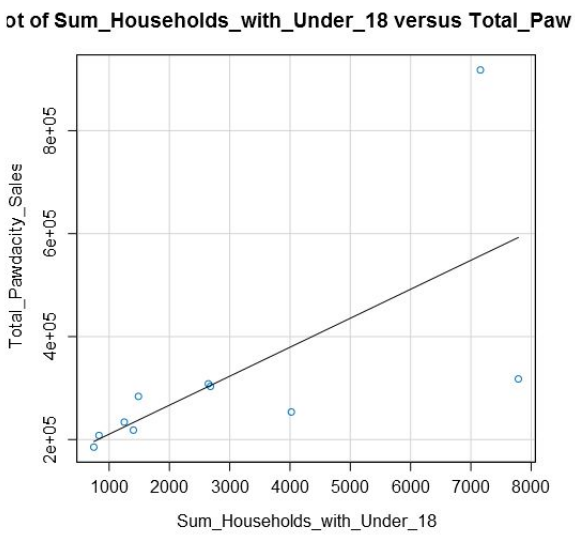
Step 1: Linear Regression

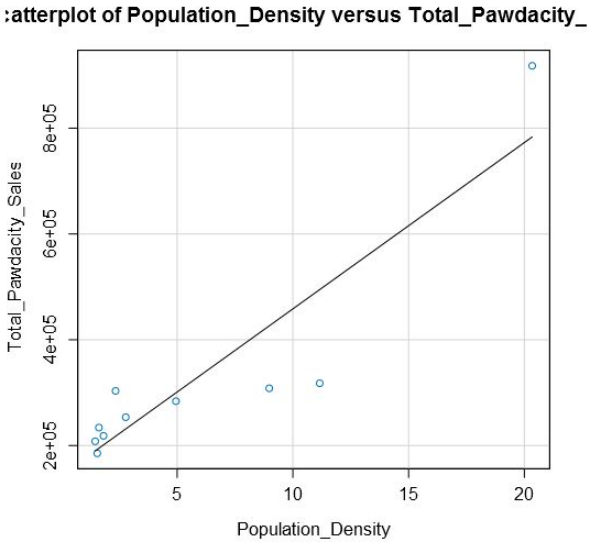
Provide an explanation of the key decisions that need to be made. (250 word limit)

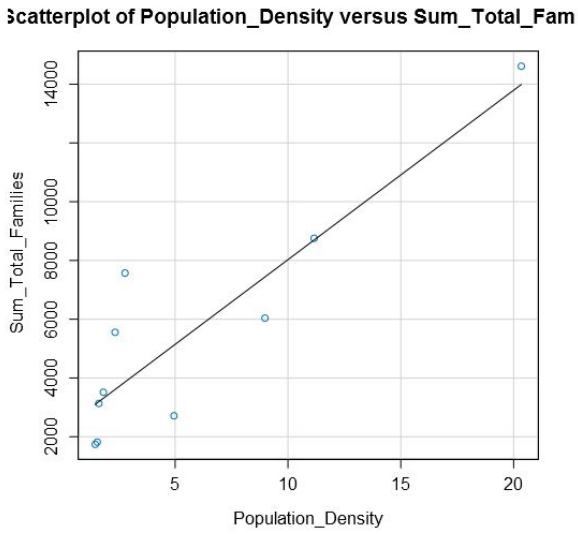
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

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

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. You must talk about 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 .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? 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 3: Analysis

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

At the minimum, answer these questions:

1. What kind of data cleaning and aggregation steps did you do?

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 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.

2. What were the sales prediction steps did you do?

I filtered my cities according to the given the criteria in the project and calculated revenue off the population density information using my linear model.

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

I would recommend the city of Laramie with a predicted sales of $305,014.