Project 2.2: Recommend a City

Note that this project is a continuation from Project 2.1: Data Cleanup. You must meet specifications for Project 2.1 before you can continue on with this Project 2.2

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.

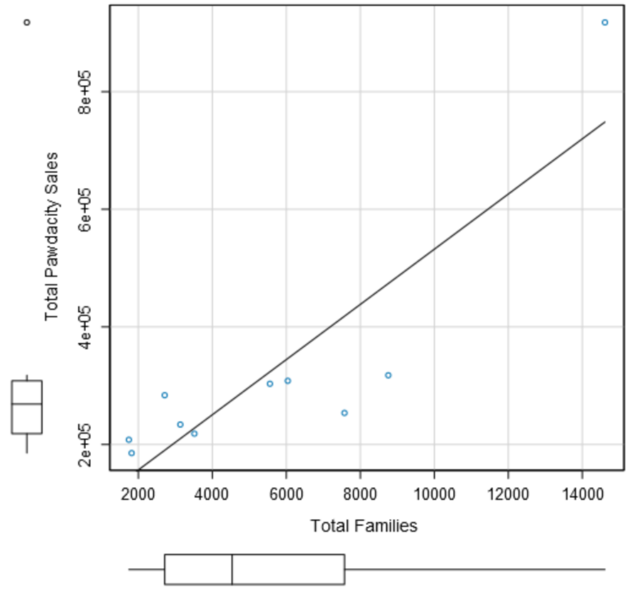
At the minimum, answer these questions:

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.**

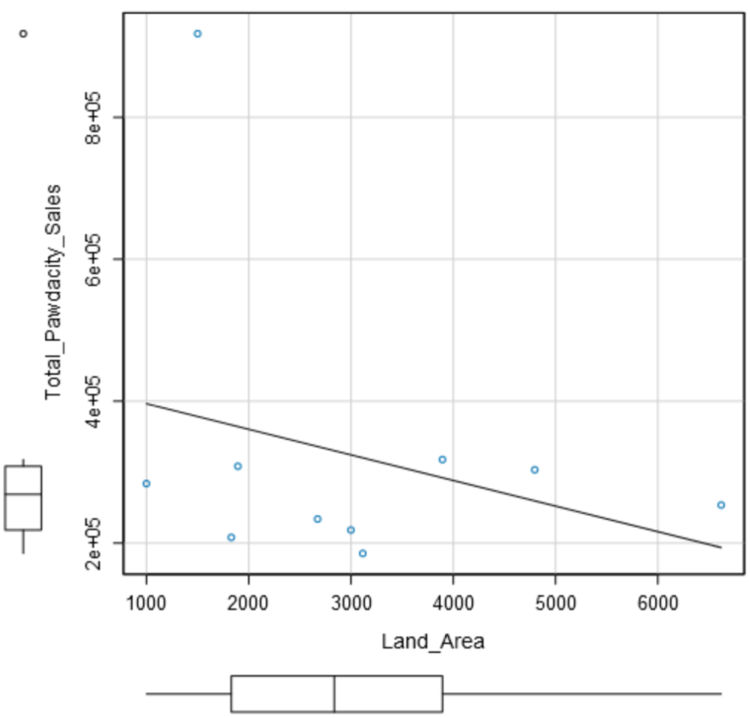
To estimate the sales amount, I choose the following predictor variables for the model

- Land Area
- Total Families

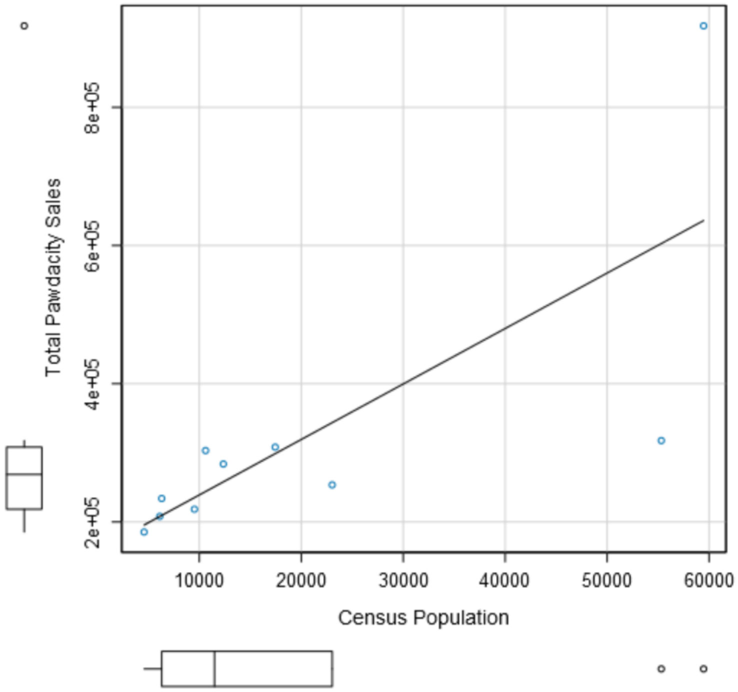
Scatter Plot of Total Pawdacity sales Vs Total families



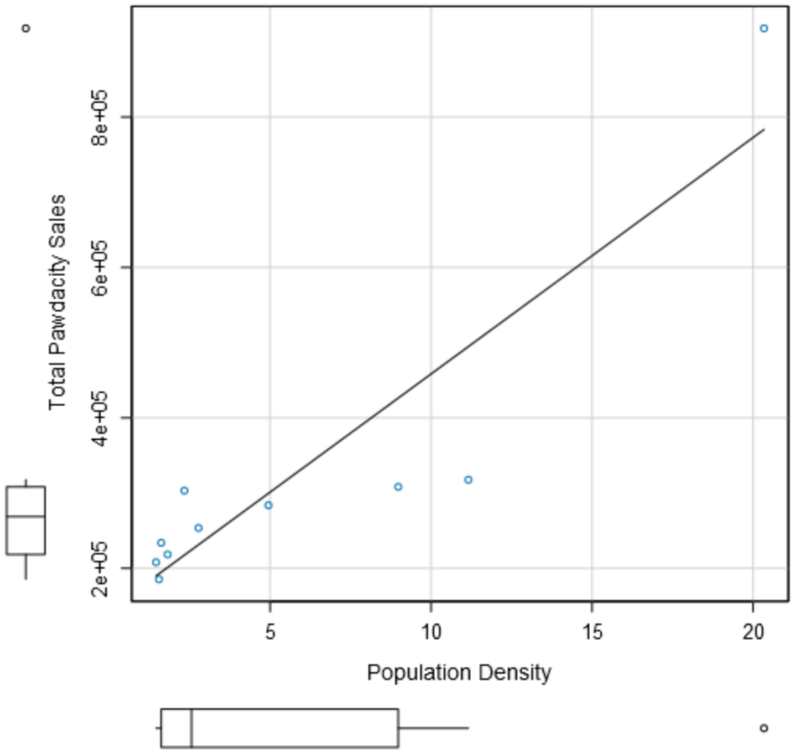
Scatter Plot of Total Pawdacity sales vs Land Area



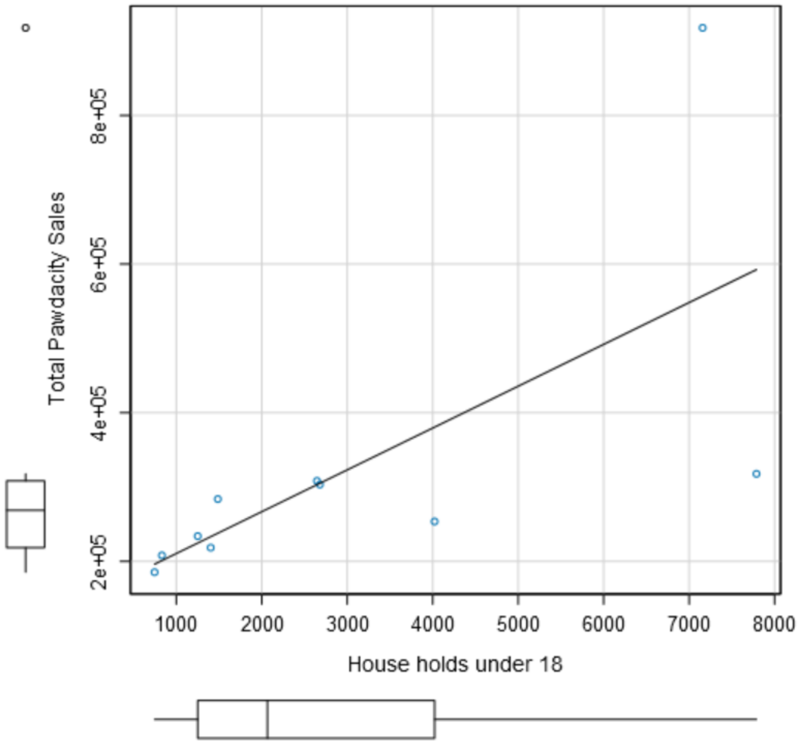
Scatter Plot of Total Pawdacity sales vs Census population



Scatter Plot of Total Pawdacity sales vs Population Density



Scatter Plot of Total Pawdacity sales vs Population Density



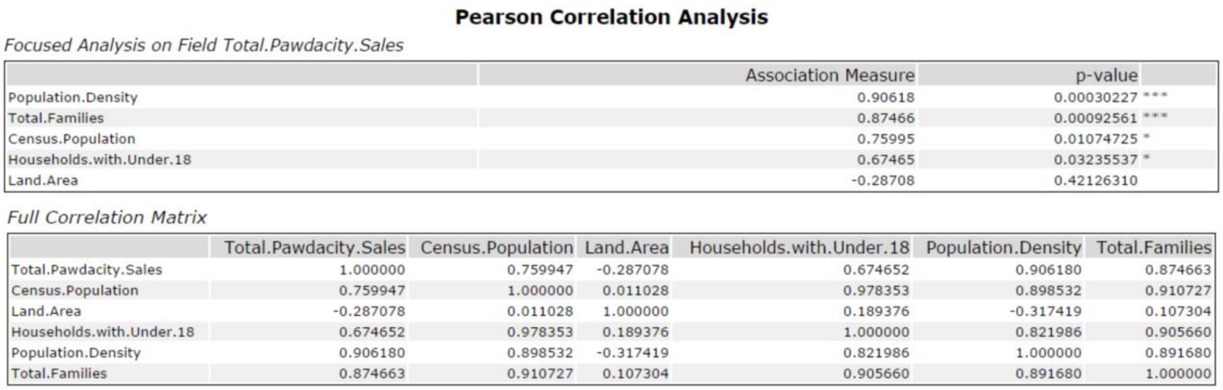
By doing the association analysis and creating a correlation matrix, I found that the Census Population, Total Families, and Population Density have strong correlations with each other, but Land are not highly correlated, So I choose the Land area as a potential predictor variable and created four models to see which of the other variables along with the Land area would result in an optimal model

The below table gives the details about the different models and resulting adjusted R squared and P values

Predictor Variables	Adjusted R squared value	P value
Land Area & Census population	0.7441	0.003519
Land Area & Households under 18	0.7081	0.005573
Land Area & Population Density	0.8771	0.00027
Land Area & Total families	0.9453	1.591e-05

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.**

When compared to the other Predictor variable combination, I see that the combination of Land area & Total families has a good Adjusted R squared and p-value, hence the model which has been created will be best suited for the predicting the sales in a new city



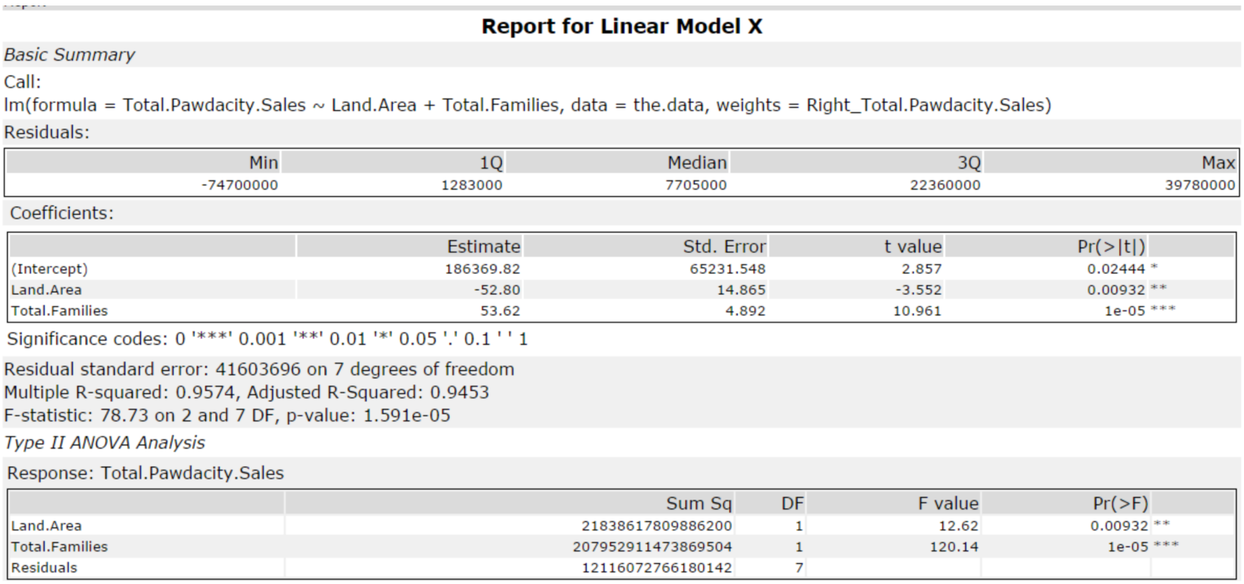
Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

	Association Measure	p-value	
Population.Density	0.90618	0.00030227	***
Total.Families	0.87466	0.00092561	***
Census.Population	0.75995	0.01074725	*
Households.with.Under.18	0.67465	0.03235537	*
Land.Area	-0.28708	0.42126310	

Full Correlation Matrix

	Total.Pawdacity.Sales	Census.Population	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Total.Pawdacity.Sales	1.000000	0.759947	-0.287078	0.674652	0.906180	0.874663
Census.Population	0.759947	1.000000	0.011028	0.978353	0.898532	0.910727
Land.Area	-0.287078	0.011028	1.000000	0.189376	-0.317419	0.107304
Households.with.Under.18	0.674652	0.978353	0.189376	1.000000	0.821986	0.905660
Population.Density	0.906180	0.898532	-0.317419	0.821986	1.000000	0.891680
Total.Families	0.874663	0.910727	0.107304	0.905660	0.891680	1.000000



Report for Linear Model X

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-74700000	1283000	7705000	22360000	39780000

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	186369.82	65231.548	2.857	0.02444	*
Land.Area	-52.80	14.865	-3.552	0.00932	**
Total.Families	53.62	4.892	10.961	1e-05	***

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

Residual standard error: 41603696 on 7 degrees of freedom
Multiple R-squared: 0.9574, Adjusted R-Squared: 0.9453
F-statistic: 78.73 on 2 and 7 DF, p-value: 1.591e-05

Type II ANOVA Analysis

Response: Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)	
Land.Area	21838617809886200	1	12.62	0.00932	**
Total.Families	207952911473869504	1	120.14	1e-05	***
Residuals	12116072766180142	7			

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)

Total Sales = 186369.82 + (-52.80 * Land Area) + (53.62 *Total Families)

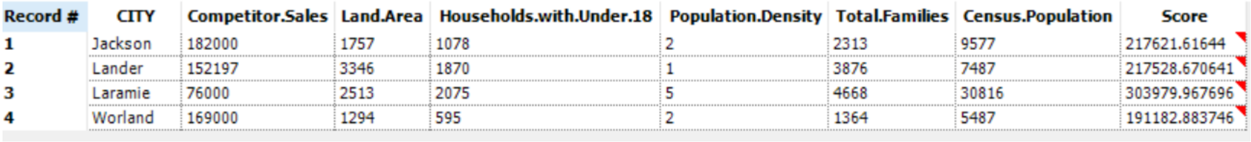
Step 2: Analysis

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

At the minimum, answer this question:

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

This is below estimated sales on each of the city where a new store could be opened

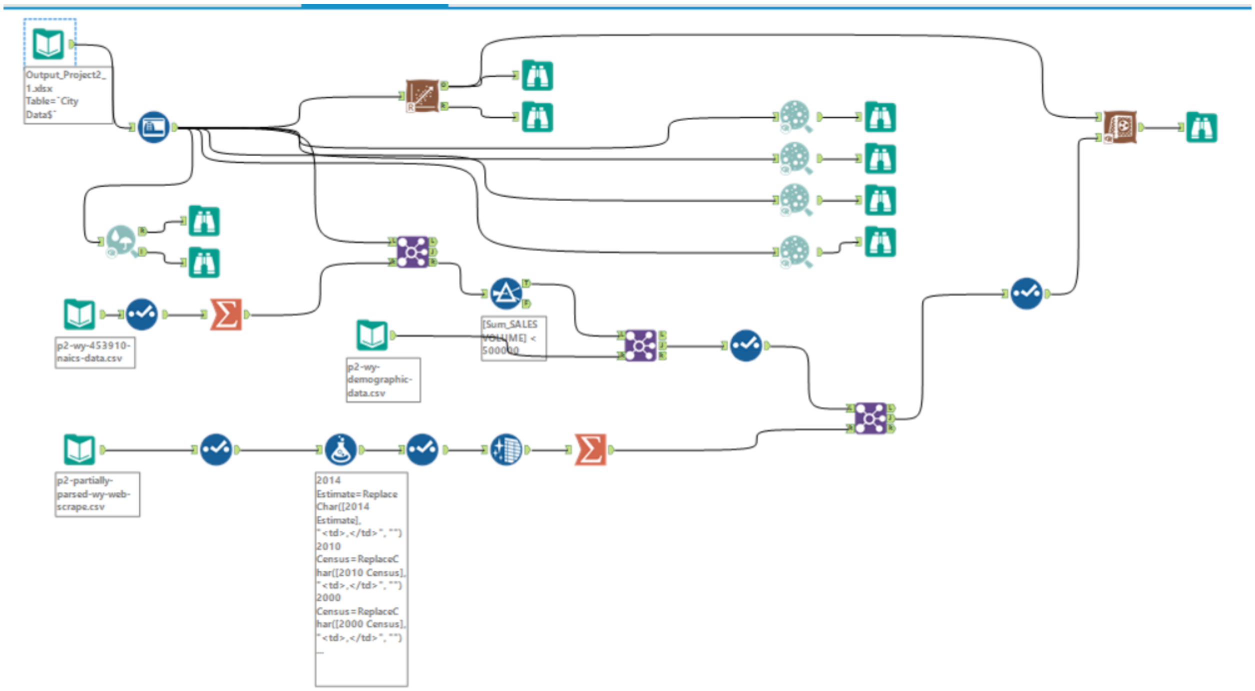


Record #	CITY	Competitor.Sales	Land.Area	Households.with.Under.18	Population.Density	Total.Families	Census.Population	Score
1	Jackson	182000	1757	1078	2	2313	9577	217621.61644
2	Lander	152197	3346	1870	1	3876	7487	217528.670641
3	Laramie	76000	2513	2075	5	4668	30816	303979.967696
4	Worland	169000	1294	595	2	1364	5487	191182.883746

From the above list, I would recommend the city of **Laramie** for the opening of the new store, The reason for that recommendation are

- Competitor sales is way too less for the particular city with the predictor variables we had chosen
- The predicted sales in the city of Laramie is $303979.97 which is more than $200,000 as given in the criteria
- The population of the city is more than 4000 according to 2014 census population
- The particular city has the highest predicted sales in the entire set of data.

Alteryx Work flow for the project


Output_Project2_1.xlsx Table=`City Data$`
p2-wy-453910-naics-data.csv
p2-wy-demographic-data.csv
[Sum_SALES VOLUME] < 500000
p2-partially-parsed-wy-web-scrape.csv
2014 Estimate=ReplaceChar([2014 Estimate], "<td>,</td>", "")
2010 Census=ReplaceChar([2010 Census], "<td>,</td>", "")
2000 Census=ReplaceChar([2000 Census], "<td>,</td>", "")
...