Project 2.2: Recommend a City

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/3d606c26-cb8e-43af-9199-7e3577aa3392/project#

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:

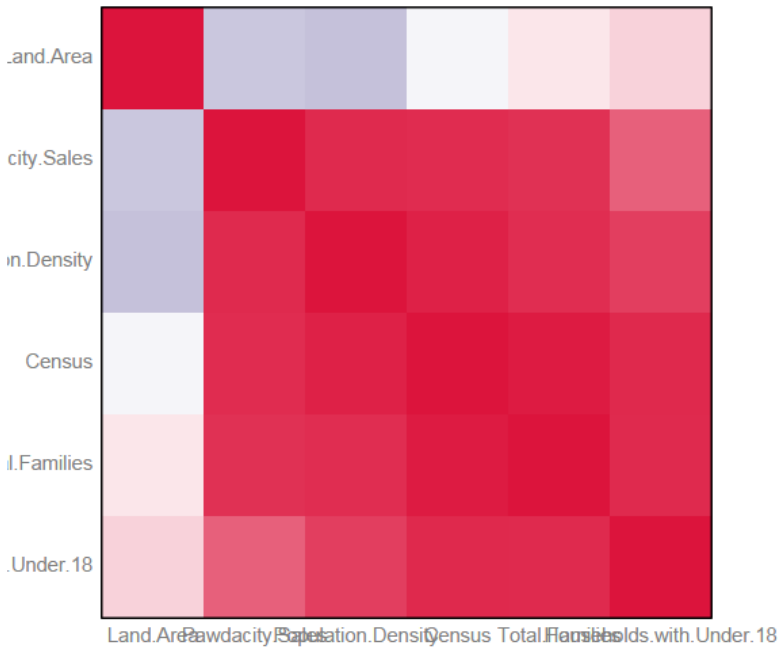
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

   When examining the correlation matrix, we can see that population related attributes are strongly correlated with each other. Census, Population Density, Households with Under 18, and Total Families have correlation values .87 or greater. These values and an understanding of what these data attributes mean suggest a strong possibility of multicollinearity.

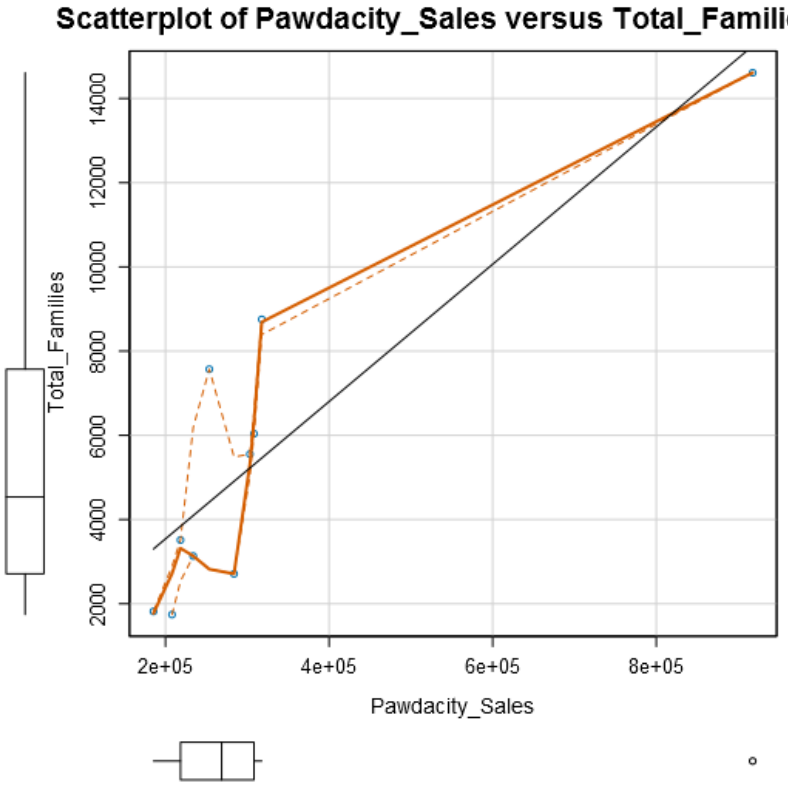
   Because Land Area is not correlated with the population values, we use it as one predictor value in our model and test to see which of the other four variables yields the best model.

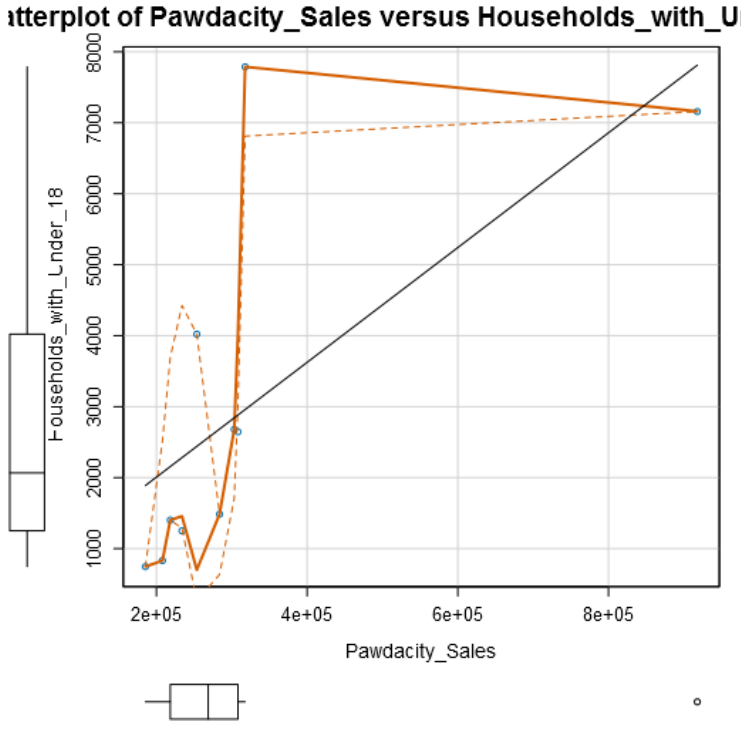
Full Correlation Matrix

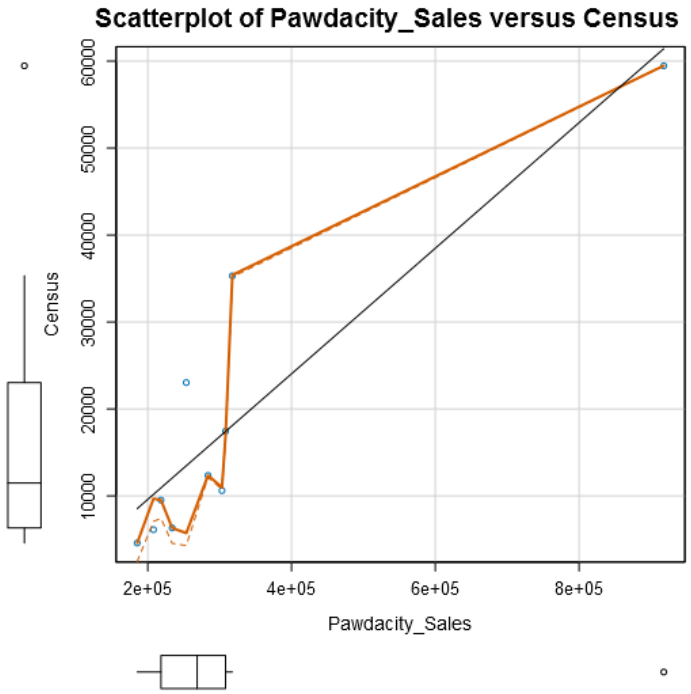
	Pawdacity.Sales	Census	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Pawdacity.Sales	1.00000	0.89875	-0.28708	0.67465	0.90618	0.87466
Census	0.89875	1.00000	-0.05247	0.91156	0.94439	0.96919
Land.Area	-0.28708	-0.05247	1.00000	0.18938	-0.31742	0.10730
Households.with.Under.18	0.67465	0.91156	0.18938	1.00000	0.82199	0.90566
Population.Density	0.90618	0.94439	-0.31742	0.82199	1.00000	0.89168
Total.Families	0.87466	0.96919	0.10730	0.90566	0.89168	1.00000

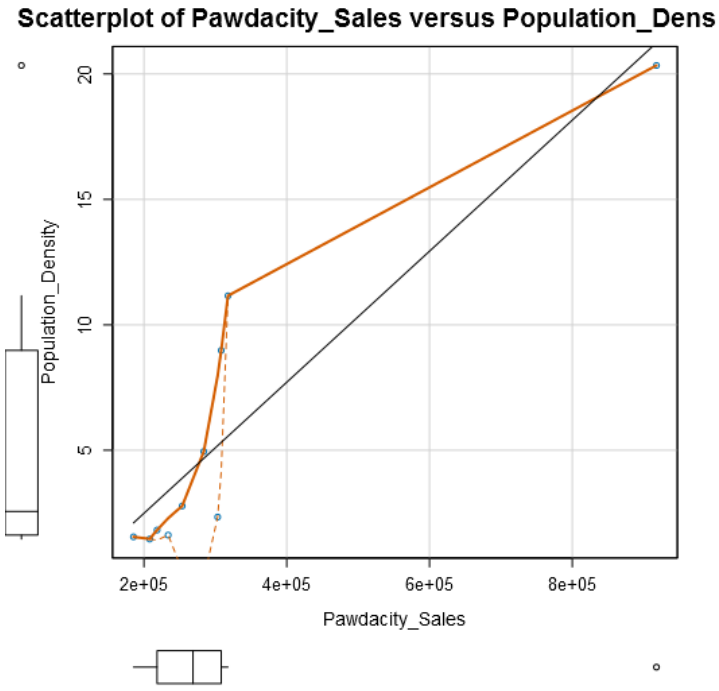

.and.Area
city.Sales
n.Density
Census
l.Families
.Under.18
Land.Area
Census
Households.with.Under.18


Scatterplot of Pawdacity_Sales versus Land_Area
Land_Area
1000
2000
3000
4000
5000
6000
2e+05
4e+05
6e+05
8e+05
Pawdacity_Sales


Scatterplot of Pawdacity_Sales versus Total_Families
Total_Families
2000
4000
6000
8000
10000
12000
14000
2e+05
4e+05
6e+05
8e+05
Pawdacity_Sales


ıtterplot of Pawdacity_Sales versus Households_with_Un
Households_with_Under_18
1000
2000
3000
4000
5000
6000
7000
8000
2e+05
4e+05
6e+05
8e+05
Pawdacity_Sales





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.

A linear model using Land Area and Total Families was the best fit. The adjusted R^2 value of .88 indicates a good fit. While the P values show a statistical significance.

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

Although a higher R^2 could be achieved by adding households under 18, p value of .07 is not statistically significant. The higher R^2 value is a result of the multicollinearity observed in the correlation matrix.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	172630.55	46790.40	3.689	0.01022	*
Land.Area	-42.81	11.69	-3.661	0.01057	*
Households.with.Under.18	-39.95	18.21	-2.194	0.07067	.
Total.Families	72.14	11.56	6.240	0.00078	***

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

Residual standard error: 57952 on 6 degrees of freedom
Multiple R-squared: 0.9511, Adjusted R-Squared: 0.9266
F-statistic: 38.89 on 3 and 6 DF, p-value: 0.0002513

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)

   Pawdacity sales = 197330.41 – (48.42 * Land Area) + (49.14 * 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?

It's my recommendation to build a new Pawdacity store in Laramie. The predicted annual sales will be $ 305,013.88.

Laramie's predicted sales exceeds the next top candidate city by approximately $79,143. This new recommended site has one competitor with annual sales of $76,000. In analysis done outside of this project, Pawdacity can tolerate competitor sales of up to $500,000.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.