

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.

Answer : I ran the Association analysis tool and opened the report option of the association analysis tool to find significance of variables to "Total Pawdacity Sales".

Alteryx Designer x64 - final Wkfiw.xml - Browse (21)

1 record displayed, 1 field, 3081 bytes

Table Report

1 of 1 Fields | Records 1 to 1

Record Layout

Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

	Association Measure	p-value
X2010.Census	0.89810	0.00017363 ***
Total.Families	0.86466	0.00059221 ***
Population.Density	0.86289	0.00062613 ***
Households.with.Under.18	0.67601	0.02239778 *

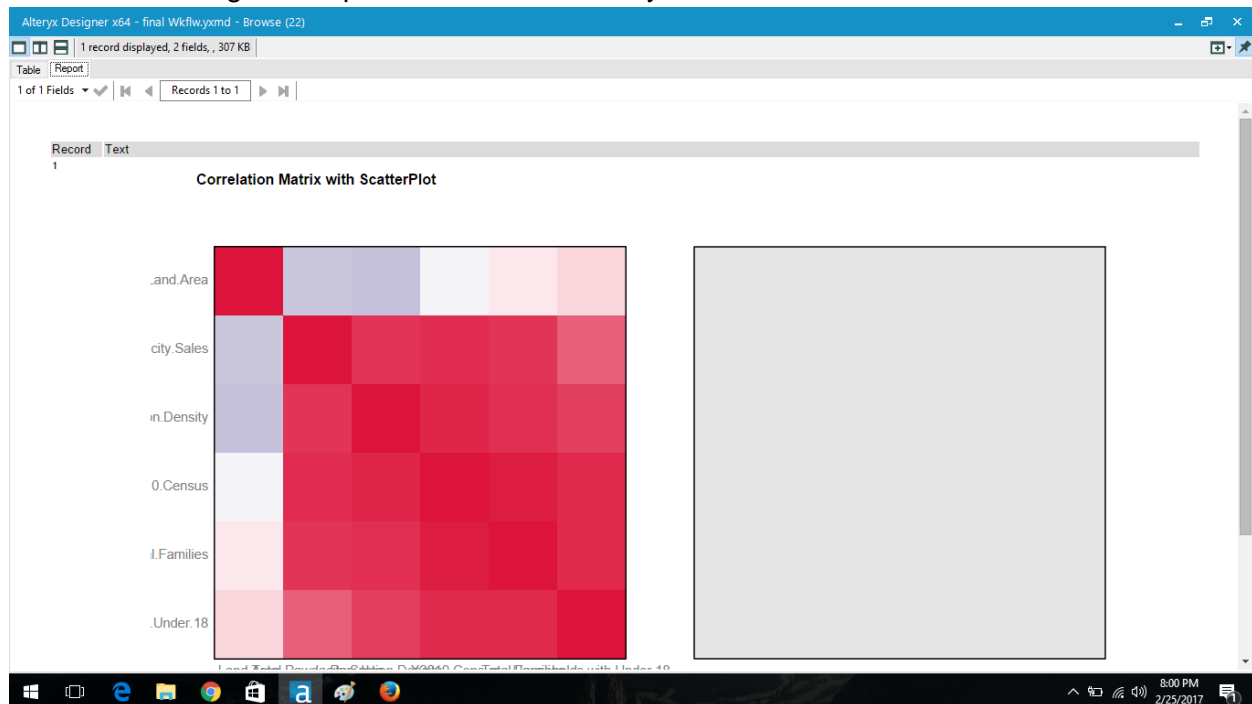
Full Correlation Matrix

	Total.Pawdacity.Sales	Households.with.Under.18	Population.Density	Total.Families	X2010.Census
Total.Pawdacity.Sales	1.00000	0.67601	0.86289	0.86466	0.89810
Households.with.Under.18	0.67601	1.00000	0.81576	0.90724	0.91188
Population.Density	0.86289	0.81576	1.00000	0.88479	0.92770
Total.Families	0.86466	0.90724	0.88479	1.00000	0.96800
X2010.Census	0.89810	0.91188	0.92770	0.96800	1.00000

Matrix of Corresponding p-values

	Total.Pawdacity.Sales	Households.with.Under.18	Population.Density	Total.Families	X2010.Census
Total.Pawdacity.Sales		2.2398e-02	6.2613e-04	5.9221e-04	1.7363e-04
Households.with.Under.18	2.2398e-02		2.2030e-03	1.1529e-04	9.2144e-05
Population.Density	6.2613e-04	2.2030e-03		2.9571e-04	3.8717e-05
Total.Families	5.9221e-04	1.1529e-04	2.9571e-04		1.0478e-06
X2010.Census	1.7363e-04	9.2144e-05	3.8717e-05	1.0478e-06	

I then tried finding similar predictor variables if any



As you can see all variables are very highly co-related(except Land area), thus there may be a presence of similar variables which would need to be excluded. Before start analyzing duplicate variables, I ran the linear regression tool, to check for any un-significant variable

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 72 KB

1 of 1 Fields

Records 1 to 10

Report for Linear Model X

Basic Summary

Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Population.Density + Total.Families + X2010.Census, data = the.data)

Residuals:

	Min	1Q	Median	3Q	Max
	-49640	-36370	-14960	23390	98670

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	290188.097	56276.523	5.1565	0.0021 **
Land.Area	-105.185	25.367	-4.1465	0.00603 **
Population.Density	-40592.150	14635.316	-2.7736	0.03227 *
Total.Families	95.330	30.932	3.0819	0.02161 *
X2010.Census	2.976	6.003	0.4957	0.63774

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

Residual standard error: 60637 on 6 degrees of freedom
Multiple R-squared: 0.9516, Adjusted R-Squared: 0.9194
F-statistic: 29.5 on 4 and 6 DF, p-value: 0.0004365

Type II ANOVA Analysis

Response: Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)
Land.Area	63218315305.53	1	17.19	0.00603 **
Population.Density	28284663500.05	1	7.69	0.03227 *
Total.Families	34922799691.9	1	9.5	0.02161 *
X2010.Census	60345128.85	1	0.25	0.63774

From the above report , we can safely remove 2010 census from further analysis as it is not significant because its p-value is above 0.05

Now to check for the best model, factoring the high co-relation between some variables

MODEL 1 : without Total Families

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 69 KB

TableReport

1 of 1 FieldsRecords 1 to 10

Record	Report																									
1	Report for Linear Model X																									
2	Basic Summary																									
3	Call: lm(formula = Total.Pawdacity.Sales ~ Land.Area + Households.with.Under.18 + Population.Density, data = the.data)																									
4	Residuals:																									
5	<table><thead><tr><th></th><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr></thead><tbody><tr><td></td><td>-166900</td><td>-37700</td><td>402</td><td>41640</td><td>213200</td></tr></tbody></table>		Min	1Q	Median	3Q	Max		-166900	-37700	402	41640	213200													
	Min	1Q	Median	3Q	Max																					
	-166900	-37700	402	41640	213200																					
6	Coefficients:																									
7	<table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr></thead><tbody><tr><td>(Intercept)</td><td>149313.325</td><td>128223.43</td><td>1.1645</td><td>0.28238</td></tr><tr><td>Land.Area</td><td>8.816</td><td>44.11</td><td>0.1998</td><td>0.84729</td></tr><tr><td>Households.with.Under.18</td><td>-14.896</td><td>47.69</td><td>-0.3123</td><td>0.76389</td></tr><tr><td>Population.Density</td><td>37368.274</td><td>20741.90</td><td>1.8016</td><td>0.11462</td></tr></tbody></table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	149313.325	128223.43	1.1645	0.28238	Land.Area	8.816	44.11	0.1998	0.84729	Households.with.Under.18	-14.896	47.69	-0.3123	0.76389	Population.Density	37368.274	20741.90	1.8016	0.11462
	Estimate	Std. Error	t value	Pr(> t)																						
(Intercept)	149313.325	128223.43	1.1645	0.28238																						
Land.Area	8.816	44.11	0.1998	0.84729																						
Households.with.Under.18	-14.896	47.69	-0.3123	0.76389																						
Population.Density	37368.274	20741.90	1.8016	0.11462																						
8	Residual standard error: 128035 on 7 degrees of freedom Multiple R-squared: 0.7483, Adjusted R-Squared: 0.6405 F-statistic: 6.939 on 3 and 7 DF, p-value: 0.01669																									
9	Type II ANOVA Analysis																									
10	Response: Total.Pawdacity.Sales																									
	<table><thead><tr><th></th><th>Sum Sq</th><th>DF</th><th>F value</th><th>Pr(>F)</th></tr></thead><tbody><tr><td>Land.Area</td><td>654653974</td><td>1</td><td>0.04</td><td>0.84729</td></tr><tr><td>Households.with.Under.18</td><td>1599130953.84</td><td>1</td><td>0.1</td><td>0.76389</td></tr><tr><td>Population.Density</td><td>53206575009.23</td><td>1</td><td>3.25</td><td>0.11462</td></tr><tr><td>Residuals</td><td>114750446523.8</td><td>7</td><td></td><td></td></tr></tbody></table>		Sum Sq	DF	F value	Pr(>F)	Land.Area	654653974	1	0.04	0.84729	Households.with.Under.18	1599130953.84	1	0.1	0.76389	Population.Density	53206575009.23	1	3.25	0.11462	Residuals	114750446523.8	7		
	Sum Sq	DF	F value	Pr(>F)																						
Land.Area	654653974	1	0.04	0.84729																						
Households.with.Under.18	1599130953.84	1	0.1	0.76389																						
Population.Density	53206575009.23	1	3.25	0.11462																						
Residuals	114750446523.8	7																								

7:08 PM
2/25/2017

MODEL 2 : without Households under 18

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 70 KB

TableReport

1 of 1 FieldsRecords 1 to 10

RecordReport

1Report for Linear Model X

2Basic Summary

3Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Population.Density + Total.Families, data = the.data)

4Residuals:

5

	Min	1Q	Median	3Q	Max
	-62880	-36800	-9538	26390	102800

6Coefficients:

7

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	287807.8	52964.22	5.434	0.00097 ***
Land.Area	-108.4	23.15	-4.685	0.00225 **
Population.Density	-39497.9	13666.17	-2.890	0.02331 *
Total.Families	106.5	19.97	5.335	0.00108 **

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

8Residual standard error: 57277 on 7 degrees of freedom
Multiple R-squared: 0.9496, Adjusted R-Squared: 0.9281
F-statistic: 44 on 3 and 7 DF, p-value: 6.542e-05

9Type II ANOVA Analysis

10Response: Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)
Land.Area	72004225021.74	1	21.95	0.00225 **
Population.Density	27403723860.43	1	8.35	0.02331 *
Total.Families	93385269123.77	1	28.47	0.00108 **
Residuals	22964308353.88	7		

7:11 PM
2/25/2017

Model 3: Without Population density

Alteryx Designer x64 - final Wkflwlyxmd - Browse (8)

12 records displayed, 2 fields, 72 KB

Table | Report

1 of 1 Fields | Records 1 to 10

Record | Report

1

Report for Linear Model X

2

Basic Summary

3

Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Households.with.Under.18 + Total.Families, data = the.data)

4

Residuals:

5

	Min	1Q	Median	3Q	Max
	-126700	-28510	6029	25050	116700

6

Coefficients:

7

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	182413.70	57216.77	3.188	0.01532 *
Land.Area	-44.67	14.33	-3.117	0.01691 *
Households.with.Under.18	-38.29	22.35	-1.713	0.13048
Total.Families	72.60	14.20	5.111	0.00138 **

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

8

Residual standard error: 71207 on 7 degrees of freedom
Multiple R-squared: 0.9222, Adjusted R-Squared: 0.8888
F-statistic: 27.64 on 3 and 7 DF, p-value: 0.0002969

9

Type II ANOVA Analysis

10

Response: Total.Pawdacity.Sales

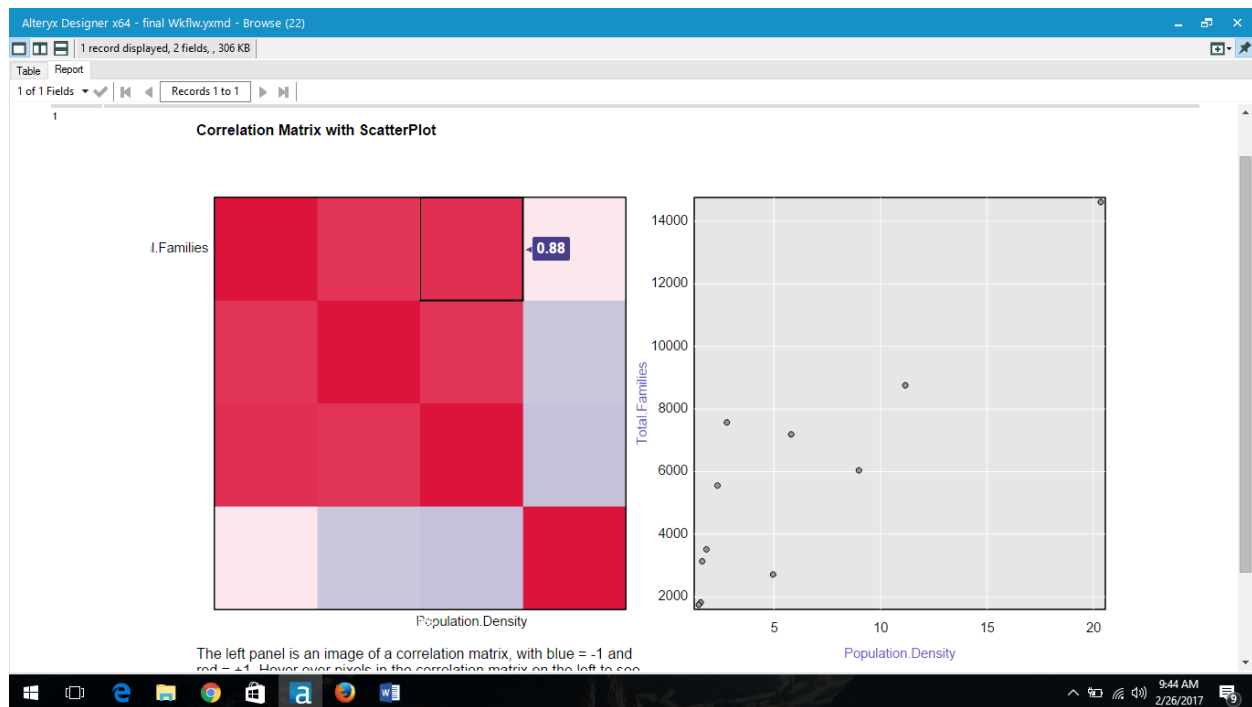
	Sum Sq	DF	F value	Pr(>F)
Land.Area	49270379646.87	1	9.72	0.01691 *
Households.with.Under.18	14875093100.76	1	2.93	0.13048
Total.Families	132464082419.49	1	26.12	0.00138 **
Residuals	35492939113.54	7		

Windows Taskbar

7:16 PM 2/25/2017

We conclude from the above the 3 models that MODEL 2 is the best fit (without Households under 18) as it has the highest Adjusted R-squared value and the lowest p-values amongst all the list of its variables.

Before we proceed with finalizing Model 2 variables, let us check for duplicate variables again with the association analysis tool:



We can see the predictor variables are still highly co-related. Let's see which variable best fits our model and eliminate the duplicate variable.

MODEL 4: Without Population density

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 69 KB

Table Report

1 of 1 Fields

Records 1 to 10

Record Report

1

Report for Linear Model X

2

3

4

5

6

7

8

9

10

Basic Summary

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

Residuals:

	Min	1Q	Median	3Q	Max
	-135400	-11930	3593	46410	111800

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	205714.49	61929.549	3.322	0.01051 *
Land.Area	-49.98	15.590	-3.206	0.0125 *
Total.Families	50.49	6.608	7.641	6e-05 ***

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

Residual standard error: 79347 on 8 degrees of freedom
Multiple R-squared: 0.8895, Adjusted R-Squared: 0.8619
F-statistic: 32.21 on 2 and 8 DF, p-value: 0.0001489

Type II ANOVA Analysis

Response: Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)
Land.Area	64706481033.6	1	10.28	0.0125 *
Total.Families	367561991716.41	1	58.38	6e-05 ***
Residuals	50368032214.31	8		

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

9:47 AM
2/26/2017

MODEL 5: without Total Families

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 71 KB

TableReport

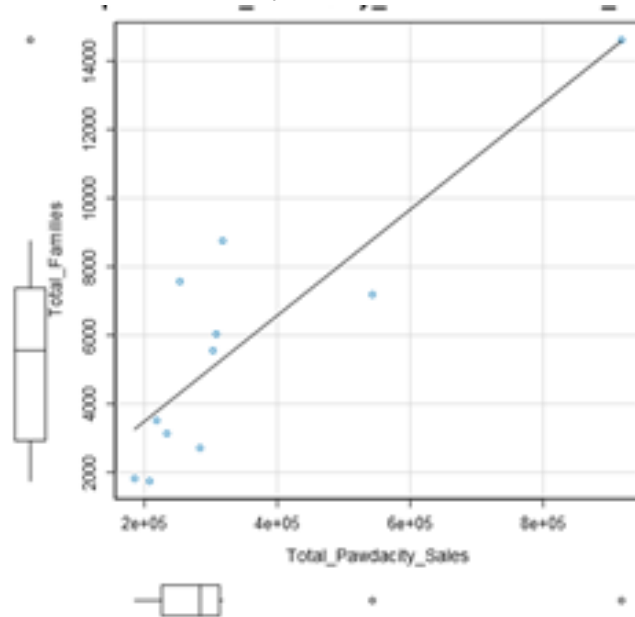
1 of 1 FieldsRecords 1 to 10

Record	Report																				
1	Report for Linear Model X																				
2	Basic Summary																				
3	Call: lm(formula = Total.Pawdacity.Sales ~ Land.Area + Population.Density, data = the.data)																				
4	Residuals:																				
5	<table><tr><td>Min</td><td>1Q</td><td>Median</td><td>3Q</td><td>Max</td></tr><tr><td>-194000</td><td>-33580</td><td>-2923</td><td>43930</td><td>196700</td></tr></table>	Min	1Q	Median	3Q	Max	-194000	-33580	-2923	43930	196700										
Min	1Q	Median	3Q	Max																	
-194000	-33580	-2923	43930	196700																	
6	Coefficients:																				
7	<table><tr><td></td><td>Estimate</td><td>Std. Error</td><td>t value</td><td>Pr(> t)</td></tr><tr><td>(Intercept)</td><td>170995.355</td><td>101543.27</td><td>1.68397</td><td>0.13068</td></tr><tr><td>Land.Area</td><td>-2.224</td><td>24.86</td><td>-0.08945</td><td>0.93092</td></tr><tr><td>Population.Density</td><td>31304.248</td><td>6874.47</td><td>4.55370</td><td>0.00187 **</td></tr></table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	170995.355	101543.27	1.68397	0.13068	Land.Area	-2.224	24.86	-0.08945	0.93092	Population.Density	31304.248	6874.47	4.55370	0.00187 **
	Estimate	Std. Error	t value	Pr(> t)																	
(Intercept)	170995.355	101543.27	1.68397	0.13068																	
Land.Area	-2.224	24.86	-0.08945	0.93092																	
Population.Density	31304.248	6874.47	4.55370	0.00187 **																	
8	Residual standard error: 120597 on 8 degrees of freedom Multiple R-squared: 0.7448, Adjusted R-Squared: 0.6811 F-statistic: 11.68 on 2 and 8 DF, p-value: 0.004239																				
9	Type II ANOVA Analysis																				
10	Response: Total.Pawdacity.Sales <table><tr><td></td><td>Sum Sq</td><td>DF</td><td>F value</td><td>Pr(>F)</td></tr><tr><td>Land.Area</td><td>116380308.5</td><td>1</td><td>0.01</td><td>0.93092</td></tr><tr><td>Population.Density</td><td>301580446453.07</td><td>1</td><td>20.74</td><td>0.00187 **</td></tr><tr><td>Residuals</td><td>116349577477.65</td><td>8</td><td></td><td></td></tr></table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Sum Sq	DF	F value	Pr(>F)	Land.Area	116380308.5	1	0.01	0.93092	Population.Density	301580446453.07	1	20.74	0.00187 **	Residuals	116349577477.65	8		
	Sum Sq	DF	F value	Pr(>F)																	
Land.Area	116380308.5	1	0.01	0.93092																	
Population.Density	301580446453.07	1	20.74	0.00187 **																	
Residuals	116349577477.65	8																			

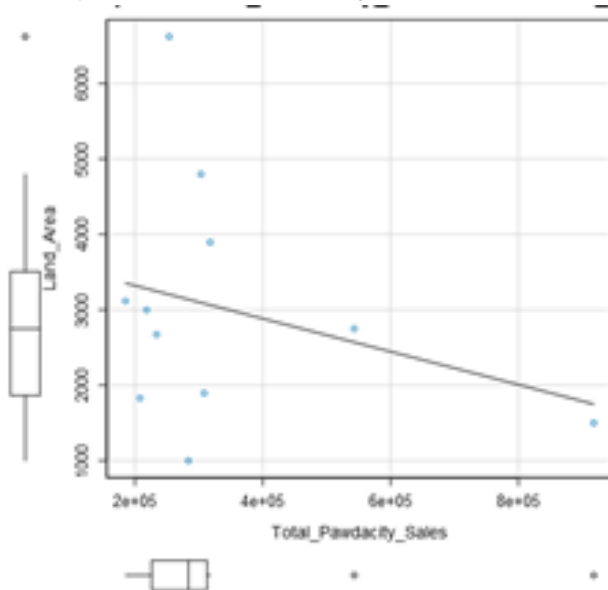
From the above two models, we can see Model 4 best fits our requirement as, the R-squared value is 0.8895 which is significantly higher and the p-values are more significant i.e. the values are less than 0.05.

And also, please find below, the scatterplots that establish linear relationship of target variable and the predictor variable (As per Model 4)

Scatterplot of Total Pawdacity sales and Total Families



Scatterplot of Total Pawdacity sales and Land area



- 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.

Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, 69 KB

Table | Report

1 of 1 Fields | Records 1 to 10

Record	Report																				
1	Report for Linear Model X																				
2	Basic Summary																				
3	Call: lm(formula = Total.Pawdacity.Sales ~ Land.Area + Total.Families, data = the.data)																				
4	Residuals:																				
5	<table><thead><tr><th></th><th>Min</th><th>1Q</th><th>Median</th><th>3Q</th><th>Max</th></tr></thead><tbody><tr><td></td><td>-135400</td><td>-11930</td><td>3593</td><td>46410</td><td>111800</td></tr></tbody></table>		Min	1Q	Median	3Q	Max		-135400	-11930	3593	46410	111800								
	Min	1Q	Median	3Q	Max																
	-135400	-11930	3593	46410	111800																
6	Coefficients:																				
7	<table><thead><tr><th></th><th>Estimate</th><th>Std. Error</th><th>t value</th><th>Pr(> t)</th></tr></thead><tbody><tr><td>(Intercept)</td><td>205714.49</td><td>61929.549</td><td>3.322</td><td>0.01051 *</td></tr><tr><td>Land.Area</td><td>-49.98</td><td>15.590</td><td>-3.206</td><td>0.0125 *</td></tr><tr><td>Total.Families</td><td>50.49</td><td>6.608</td><td>7.641</td><td>6e-05 ***</td></tr></tbody></table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Estimate	Std. Error	t value	Pr(> t)	(Intercept)	205714.49	61929.549	3.322	0.01051 *	Land.Area	-49.98	15.590	-3.206	0.0125 *	Total.Families	50.49	6.608	7.641	6e-05 ***
	Estimate	Std. Error	t value	Pr(> t)																	
(Intercept)	205714.49	61929.549	3.322	0.01051 *																	
Land.Area	-49.98	15.590	-3.206	0.0125 *																	
Total.Families	50.49	6.608	7.641	6e-05 ***																	
8	Residual standard error: 79347 on 8 degrees of freedom Multiple R-squared: 0.8895, Adjusted R-Squared: 0.8619 F-statistic: 32.21 on 2 and 8 DF, p-value: 0.0001489																				
9	Type II ANOVA Analysis																				
10	Response: Total.Pawdacity.Sales																				
	<table><thead><tr><th></th><th>Sum Sq</th><th>DF</th><th>F value</th><th>Pr(>F)</th></tr></thead><tbody><tr><td>Land.Area</td><td>64706481033.6</td><td>1</td><td>10.28</td><td>0.0125 *</td></tr><tr><td>Total.Families</td><td>367561991716.41</td><td>1</td><td>58.38</td><td>6e-05 ***</td></tr><tr><td>Residuals</td><td>50368032214.31</td><td>8</td><td></td><td></td></tr></tbody></table> <p>Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</p>		Sum Sq	DF	F value	Pr(>F)	Land.Area	64706481033.6	1	10.28	0.0125 *	Total.Families	367561991716.41	1	58.38	6e-05 ***	Residuals	50368032214.31	8		
	Sum Sq	DF	F value	Pr(>F)																	
Land.Area	64706481033.6	1	10.28	0.0125 *																	
Total.Families	367561991716.41	1	58.38	6e-05 ***																	
Residuals	50368032214.31	8																			

The linear model is a good model as all the predictor variables are highly significant (P-values are less than 0.05)

And also, the adjusted R-squared value is 0.8619 which means nearly all variance is covered by the model and is well above the standard of 0.50 used in practice.

- 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)

ANSWER : $Y = m + b_0x_0 + b_1x_1 + b_2x_2$

Total pawdacity sales = Intercept + (-49.98*Land area) + (50.49*Total families)

Total pawdacity sales =205714.49+ (-49.98*Land area) + (50.49*Total families)

Step 2: Analysis

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

At the minimum, answer this question:

- Which city would you recommend and why did you recommend this city?

I would recommend the 14th store to be opened on the city of "Laramie" as according to my calculations, it would generate the highest sales considering the given conditions/criteria of the project

The sales generated is estimated at \$315,812.7370

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.