Recomende uma cidade

Observe que este projeto é uma continuação do projeto de limpeza de dados.

Passo 1: Regressão Linear

Criando o Modelo

Abaixo está o conjunto de dados final usado para modelo de regressão.

City	Census_Population	Household_With_Under_18	Land_Area	Padacity_Sales	Population_Density	Total_Families
Buffalo	4585	746	3115.5075	185328	1.55	1819.5
Casper	35316	7788	3894.3091	317736	11.16	8756.32
Cheyenne	59466	7158	1500.1784	917892	20.34	14612.64
Cody	9520	1403	2998.95696	218376	1.82	3515.62
Douglas	6120	832	1829.4651	208008	1.46	1744.08
Evanston	12359	1486	999.4971	283824	4.95	2712.64
Powell	6314	1251	2673.57455	233928	1.62	3134.18
Riverton	10615	2680	4796.859815	303264	2.34	5556.49
Rock Springs	23036	4022	6620.201916	253584	2.78	7572.18
Sheridan	17444	2646	1893.977048	308232	8.98	6039.71

Selecionando as vereáveis preditoras.

Abaixo está a tabela de correlação Pearson de todas as variáveis.

Pearson Correlation Analysis

Focused Analysis on Field Total_Pawdacity_Sales

	Association Measure	p-value
Population_Density	0.90618	0.00030227***
X2010_Census	0.89875	0.00040617***
Total_Families	0.87466	0.00092561***
Households_With_Under_18	0.67465	0.03235536*
Land_Area	-0.28708	0.42126309

Matrix completa de correlação.

Full Correlation Matrix

	Total_ Pawdacity_ Sales	X2010_ Census	Land_ Area	Households_ With_Under_18	Population_ Density	Total_ Families
Total_ Pawdacity_Sales	1.00000	0.89875	-0.28708	0.67465	0.90618	0.87466
X2010_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

Matrix do p-valores para as variáveis preditoras.

Matrix of Corresponding p-values

	Total_ Pawdacity_ Sales	X2010_ Census	Land_ Area	Households_ With_Under_18	Population_ Density	Total_ Families
Total_ Pawdacity_Sales		4.0617e-04	4.2126e-01	3.2355e-02	3.0227e-04	9.2561e-04
X2010_Census	4.0617e-04		8.8554e-01	2.4026e-04	3.9116e-05	3.7983e-06
Land_Area	4.2126e-01	8.8554e-01		6.0028e-01	3.7148e-01	7.6796e-01
Households_ With_Under_18	3.2355e-02	2.4026e-04	6.0028e-01		3.5227e-03	3.0884e-04
Population_ Density	3.0227e-04	3.9116e-05	3.7148e-01	3.5227e-03		5.2748e-04
Total_Families	9.2561e-04	3.7983e-06	7.6796e-01	3.0884e-04	5.2748e-04	

A matriz de correlação mostra boa correlação entre as variáveis preditivas, 2010_Census, Census_Population, Households_with_Under_18, Population_Density e Total_Families. Podendo haver alguma multicolinearidade.

A Land_Area não mostra grande correlação com as outras variáveis preditivas, portanto, começarei executando uma regressão linear com Land_Area e ir adicionando as outras variáveis preditoras à regressão.

1. New_Store_LM - 1_Sales_VS_Land

Report for Linear Model New_Store_LM___1_Sales_VS_Land

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ Land_Area, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-158398	-110852	-78938	39381	539607

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	432401.73	146208.93	2.9574	0.01822*
Land_Area	-36.07	42.56	-0.8477	0.42126

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

Erro padrão residual: 217368 em 8 graus de liberdade

R quadrada múltipla: 0.08241, R quadrada ajustada: -0.03228

F estatístico: 0.7185 em 1 e 8 graus de liberdade (DF), valor p 0.4213

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	33949588837.33	1	0.72	0.42126
Residuals	377992295324.27	8		

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

Basic Diagnostic Plots

R quadrada do modelo linear para Total_Pawdacity_Sales X Land_Area = **0.08241**

2. New_Store_LM - 2_Sales_VS_Land_VS_Census

Report for Linear Model New_Store_LM___2_Sales_VS_Land_VS_Census

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ X2010_Census + Land_Area, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-164955	-28633	-9045	30193	120319

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	210872.04	69180.625	3.048	0.01863*
X2010_Census	11.03	1.728	6.383	0.00037***
Land_Area	-30.23	17.443	-1.733	0.12668

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

Erro padrão residual: 88974 em 7 graus de liberdade

R quadrada múltipla: 0.8655, R quadrada ajustada: 0.827

F estatístico: 22.52 em 2 e 7 graus de liberdade (DF), valor p 0.0008928

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
X2010_Census	322578046861.07	1	40.75	0.00037***
Land_Area	23777499407.94	1	3	0.12668
Residuals	55414248463.21	7		

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

Basic Diagnostic Plots

R quadrada ajustada do modelo linear para Total_Pawdacity_Sales X Land_Area X 2010_Census = **0.827**

3. New_Store_LM - 3_Sales_VS_Land_VS_Households_w/u_18

Report for Linear Model New_Store_LM___3_Sales_VS_Land_VS_Households_w_u_18

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ Land_Area + Households_With_Under_18, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-260680	-50922	-1834	47389	249780

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	297611.68	107140.63	2.778	0.02739*
Land_Area	-54.07	29.28	-1.847	0.10727
Households_With_Under_18	63.09	19.44	3.245	0.01415*

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

Erro padrão residual: 146831 em 7 graus de liberdade

R quadrada múltipla: 0.6336, R quadrada ajustada: 0.529

F estatístico: 6.054 em 2 e 7 graus de liberdade (DF), valor p 0.02976

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	73529107680.71	1	3.41	0.10727
Households_With_Under_18	227077058908.59	1	10.53	0.01415*
Residuals	150915236415.68	7		

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

Basic Diagnostic Plots

R quadrada ajustada do modelo linear para Total_Pawdacity_Sales X Land_Area X Housesholdes_W/u_18 = **0.529**

4. New_Store_LM - 4_Sales_VS_Land_VS_Pop_Density

Report for Linear Model New_Store_LM___4_Sales_VS_Land_VS_Pop_Density

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ Land_Area + Population_Density, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-177058	-13382	17904	34966	134588

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.435e+05	87450.27	1.641189	0.14476
Land_Area	7.846e-02	21.18	0.003704	0.99715
Population_Density	3.145e+04	5848.33	5.377362	0.00103**

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

Erro padrão residual: 102588 em 7 graus de liberdade

R quadrada múltipla: 0.8212, R quadrada ajustada: 0.7701

F estatístico: 16.07 em 2 e 7 graus de liberdade (DF), valor p 0.002419

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	144414.54	1	0	0.99715
Population_Density	304321939965.4	1	28.92	0.00103**
Residuals	73670355358.88	7		

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

Basic Diagnostic Plots

R quadrada ajustado do modelo linear para Total_Pawdacity_Sales X Land_Area X Population_Density = **0.7701**

5. New_Store_LM - 5_Sales_VS_Land_VS_T_Families

Report for Linear Model New_Store_LM___5_Sales_VS_Land_VS_T_Families

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121261	-4453	8418	40491	75205

Coeficientes:

	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

Erro padrão residual: 72030 em 7 graus de liberdade

R quadrada múltipla: 0.9118, R quadrada ajustada: 0.8866

F estatístico: 36.2 em 2 e 7 graus de liberdade (DF), valor p 0.0002035

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	60473052720.43	1	11.66	0.01123*
Total_Families	341673845917.83	1	65.85	8e-05***
Residuals	36318449406.44	7		

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

Basic Diagnostic Plots

R quadrada ajustado do modelo linear para Total_Pawdacity_Sales X Land_Area X Total_Families = **0.8866**

6. New_Store_LM - 6_Sales_VS_Land_VS_T_Families_VS_Census.

Report for Linear Model New_Store_LM___6_Sales_VS_Land_VS_T_ Families_VS_Census

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ X2010_Census + Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-110035	-4750	10184	41556	75241

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	196536.22	60172.001	3.2662	0.01711*
X2010_Census	-3.21	7.855	-0.4087	0.69697
Land_Area	-53.55	19.644	-2.7262	0.03436*
Total_Families	62.78	33.998	1.8465	0.11434

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

Erro padrão residual: 76741 em 6 graus de liberdade

R quadrada múltipla: 0.9142, R quadrada ajustada: 0.8713

F estatístico: 21.32 em 3 e 6 graus de liberdade (DF), valor p 0.001335

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

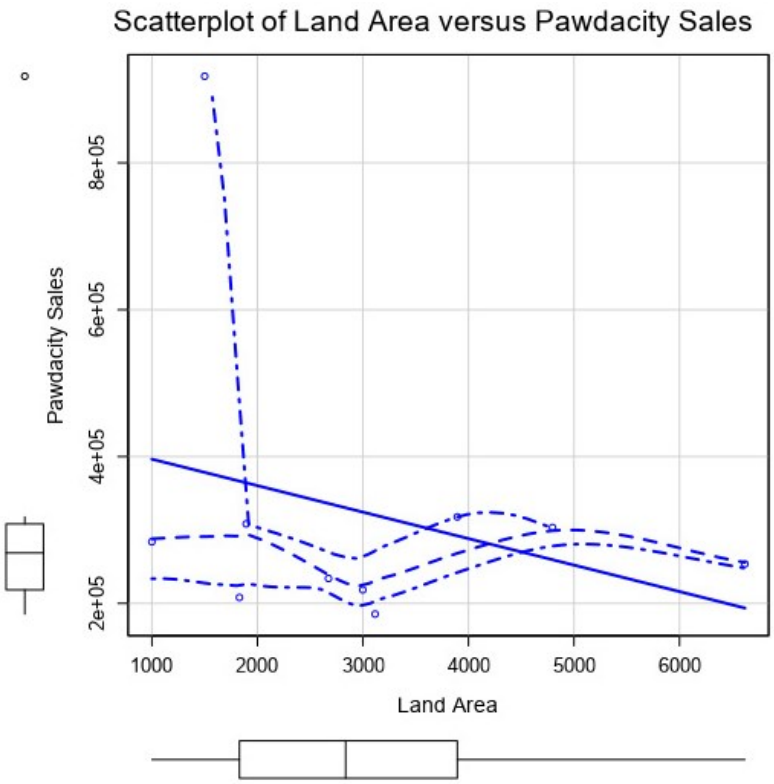
	Sum Sq	DF	F value	Pr(>F)
X2010_Census	983564136.27	1	0.17	0.69697
Land_Area	43768907210.74	1	7.43	0.03436*
Total_Families	20079363193.04	1	3.41	0.11434
Residuals	35334885270.17	6		

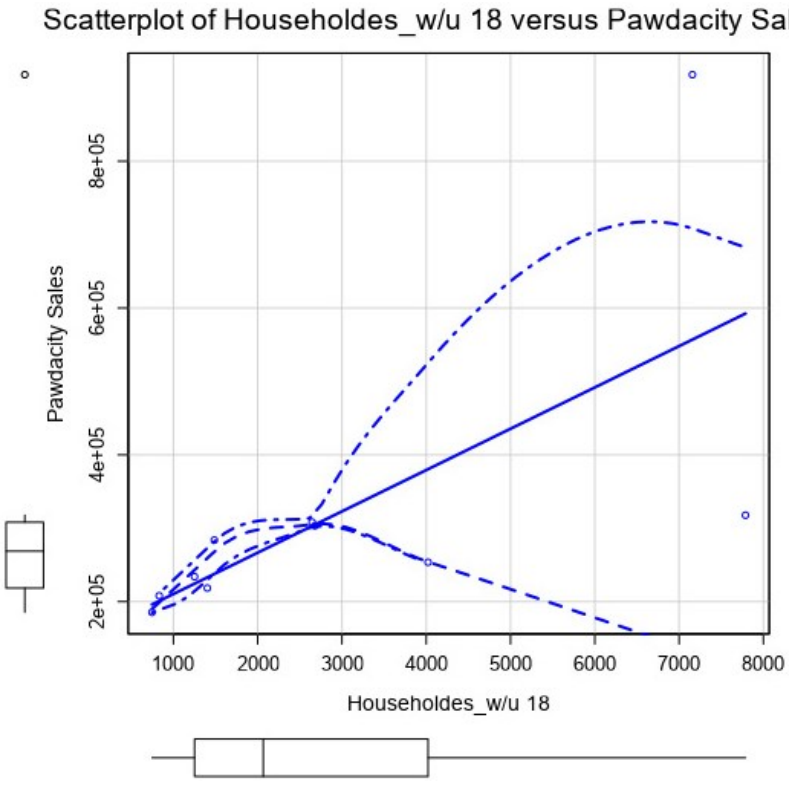
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

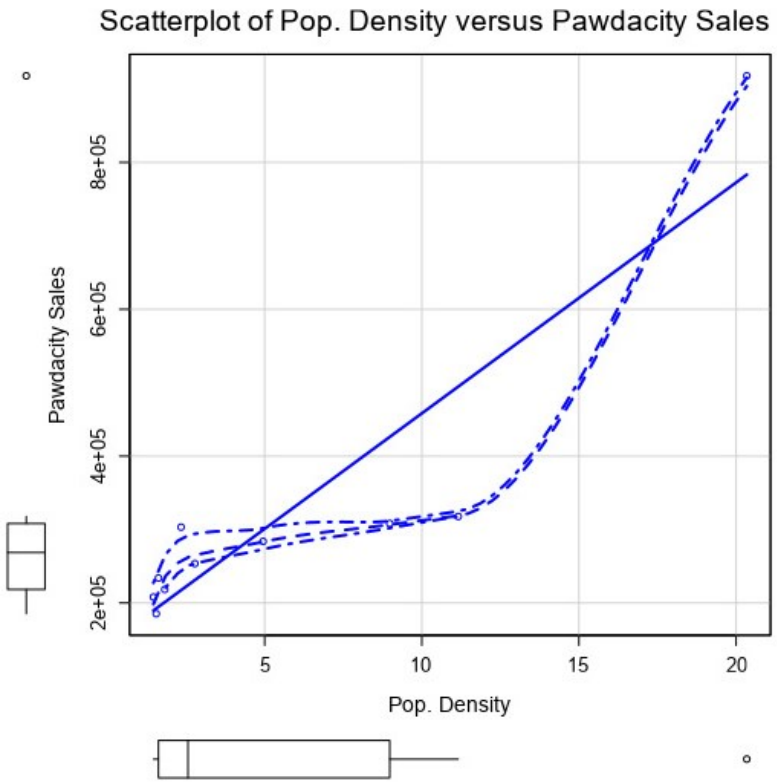
Basic Diagnostic Plots

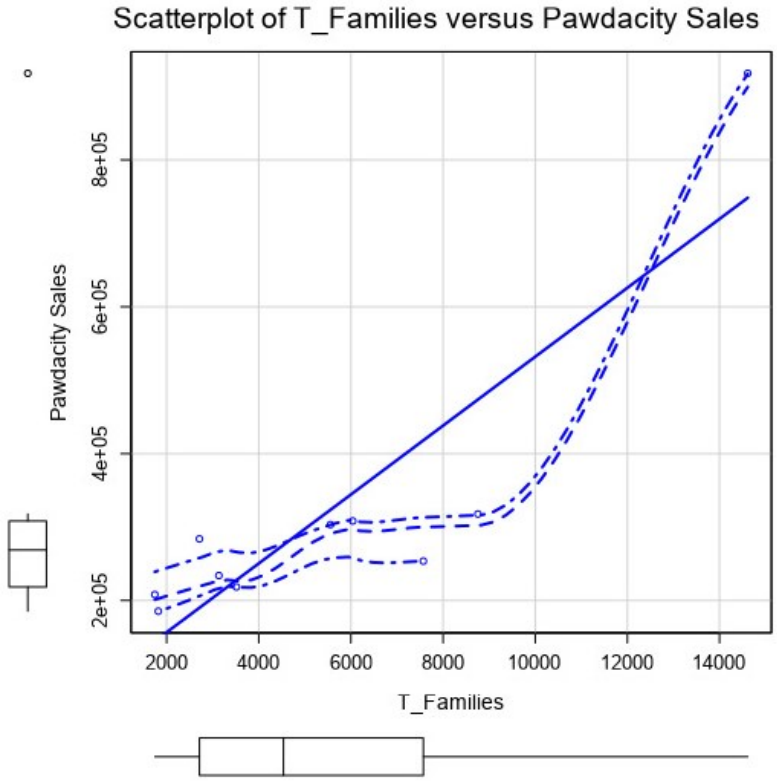
R quadrada ajustado do modelo linear para Total_Pawdacity_Sales X Land_Area X Total_Families X 2010 Census = **0.8713**

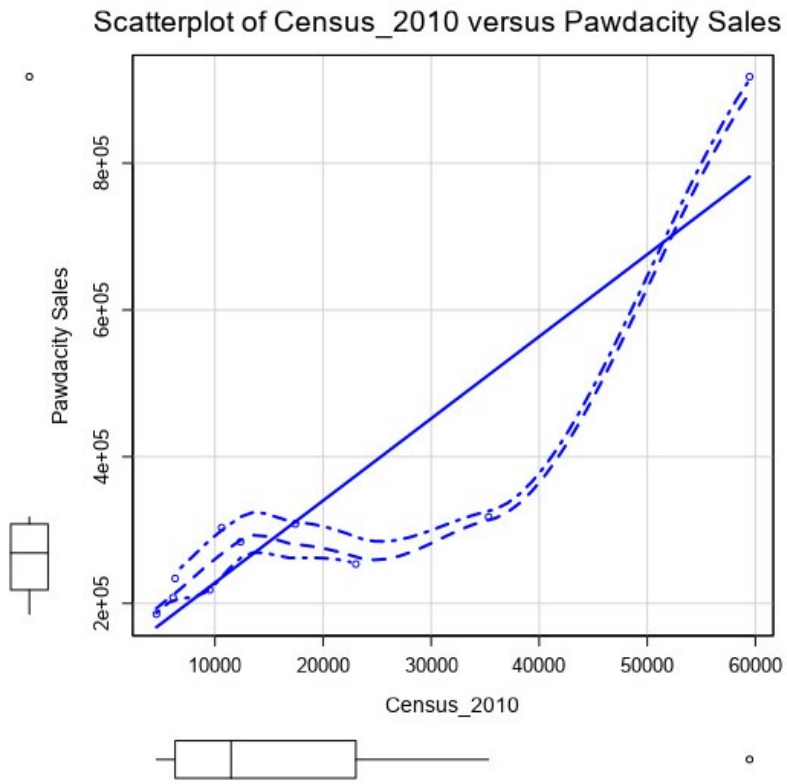
Abaixo estão os Scatterplot para todas as variáveis preditoras X variável alvo (Total_Pawdacity_Sales).






Scatterplot of Pop. Density versus Pawdacity Sales
Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Pop. Density


Scatterplot of T_Families versus Pawdacity Sales
Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
2000
4000
6000
8000
10000
12000
14000
T_Families



Os gráficos de dispersão acima fornecem uma boa representação da linearidade entre a variável alvo (Total_Pawdacity_sales) e sua respectiva variável preditora.

Começando com Land_Area como uma variável preditora (R-quadrado = 0,08241) e adicionando outras variáveis, possível ver uma maior diferença no R-quadrado quando estão sendo usadas as variáveis Land_Area e Total_Families (R-quadrado ajustado = 0,8866).

Usarei Land_Area e Total_Families como minhas variáveis preditoras para o meu modelo linear.

Abaixo o resumo do modelo de regressão multilinear.

Report for Linear Model New_Store_LM___5_Sales_VS_Land_VS_T_ Families

Basic Summary

Call:

lm(formula = Total_Pawdacity_Sales ~ Land_Area + Total_Families, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-121261	-4453	8418	40491	75205

Coeficientes:

	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

Erro padrão residual: 72030 em 7 graus de liberdade

R quadrada múltipla: 0.9118, R quadrada ajustada: 0.8866

F estatístico: 36.2 em 2 e 7 graus de liberdade (DF), valor p 0.0002035

Type II ANOVA Analysis

Response: Total_Pawdacity_Sales

	Sum Sq	DF	F value	Pr(>F)
Land_Area	60473052720.43	1	11.66	0.01123*
Total_Families	341673845917.83	1	65.85	8e-05***
Residuals	36318449406.44	7		

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

Basic Diagnostic Plots

A equação do modelo de regressão linear é:

Y (Total_Pawdacity_Sales) = 1973330.41 – 48.42 (Land_area) + 49.12 (Total_Families).

Análise – Recomendação.

Esses são os critérios para a escolha da nova cidade:

1. A nova loja deve estar localizada em uma nova cidade. Isso significa que não deve haver lojas existentes na nova cidade.
2. O total de vendas para toda a competição na nova cidade deve ser inferior a US $ 500.000
3. A nova cidade onde você deseja construir sua nova loja deve ter uma população superior a 4.000 pessoas (com base na estimativa do Censo dos EUA em 2014).
4. As vendas anuais previstas devem ser superiores a US $ 200.000.
5. A cidade escolhida tem as vendas previstas mais altas do conjunto previsto.

Com os critérios acima, recomendo a Laramie City, que atualmente não contém uma loja e tem uma população estimada em 2014 de 32.081 habitantes e previsão de vendas é de US $ 305.013,88.

Abaixo esta um resumo das 6 possíveis cidades para abertura da 14° loja, em destaque está minha recomendação.

City	2014_Census	Total_Families	Score_Pawdacity_sales
Laramie	32081.00	4668.93	305013.88
Torrington	6736.00	2548.50	245081.79
Jackson	10449.00	2313.08	225870.82
Lander	7642.00	3876.81	225751.40
Green River	12630.00	3977.40	224372.00
Worland	5366.00	1364.32	201700.33

Alterxy Workflow