Project 2.1: Data Cleanup

Passo 1: Entendimento do Negócio e dos

1. Que decisões devem ser tomadas?

Temos a empresa Pawdacity, a qual e líder no Mercado de pets no estado de Wyoming, com 13 lojas distribuídas em todo o estado, o nosso objetivo e criar um modelo preditivo para ajudar na tomada de decisão da empresa ao abrir a 14 loja, para lograr o objetivo e necessário analisar dados de clientes, competência, e dados demográficos do estado.

2. Que dados são necessários para subsidiar essas decisões?

E necessário analisar o numero de potenciais clientes, os dados de consumo, os dados das lojas da competência, e as características demográficas das regiões do estado.

Passo 2: Construindo o Conjunto de Treinamento

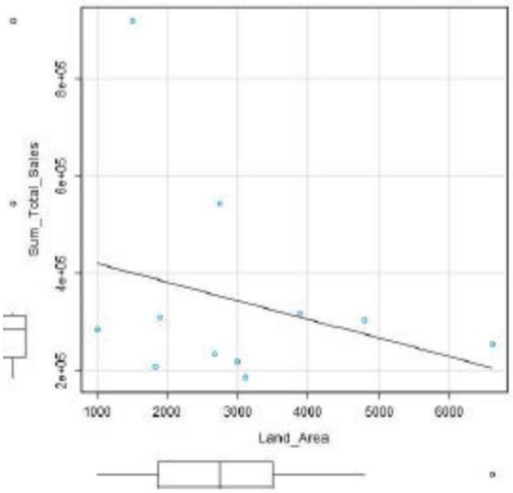
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Passo 3: Tratando os Outliers

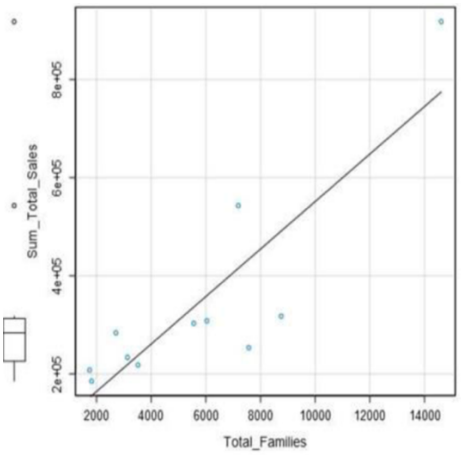
Existem cidades que são outliers no conjunto de treinamento? Qual outlier você escolheu para remover ou imputar? Como esse conjunto de dados é um conjunto de dados pequeno (11 cidades), **você deve apenas remover ou imputar um outlier**. Explique o seu raciocínio.

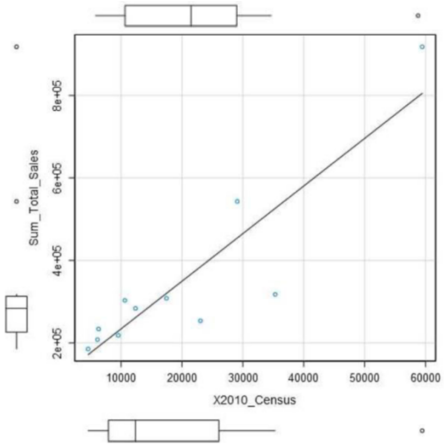
	A	B	C	D	E	F	G
1	CITY	Sum_Total Sales	Land Area	ouseholds with Under	Population Density	Total Families	2010 Census
2	Buffalo	-	-	-	-	-	-
3	Casper	-	-	-	-	-	-
4	Cheyenne	917.892	-	-	20,34	14.612,64	59.466
5	Cody	-	-	-	-	-	-
6	Douglas	-	-	-	-	-	-
7	Evanston	-	-	-	-	-	-
8	Gillette	543.132	-	-	-	-	-
9	Powell	-	-	-	-	-	-
10	Riverton	-	-	-	-	-	-
11	Rock Springs	-	6.620,20	-	-	-	-
12	Sheridan	-	-	-	-	-	-
13							
14	Q1	226152,00	1861,72	1327,00	1,72	2923,41	7917,00
15	Q3	312984,00	3504,91	4037,00	7,39	7380,81	26061,50
16	Q3 - Q1	86832,00	1643,19	2710,00	5,67	4457,40	18144,50
17	lower	95904,00	-603,06	-2738,00	-6,78	-3762,68	-19299,75
18	upper	443232,00	5969,69	8102,00	15,89	14066,90	53278,25

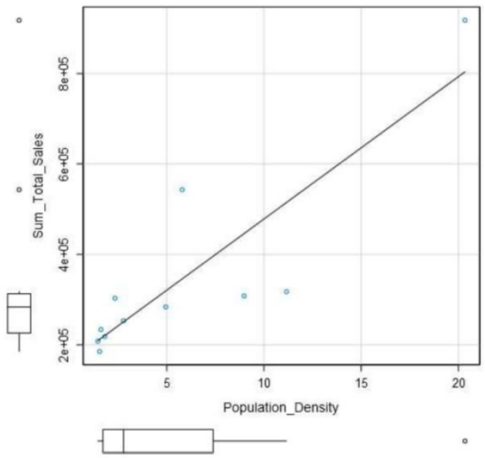
A cidade Rock Springs e um outlier na coluna “Land_Area”, em este caso não precisamos remover o dado, já que não interfere na curva de relacionamento.



A cidade Cheyenne tem três outliers, como podemos observar nos scatterplot seguintes.







Não devemos remover estes outliers devido a nosso conjunto já ser pequeno e os outliers não interferir na curva de relacionamento.

No entanto, na cidade de Gillette, podemos apreciar um outlier na coluna “Sum_Total_Sales”, esta coluna sobre valora a media de vendas nas cidades pequenas principalmente, por tanto devemos remover sim este outlier.