Projeto 1: Prevendo Demanda de um Catálogo

Complete cada seção. Quando estiver pronto, salve o arquivo como um documento PDF e envie-o aqui: https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

Passo 1: Compreensão do Negócio e dos Dados

De forma bem resumida, o projeto tem como objetivo saber se é possível ter um lucro maior do que U$ 10.000,00 ao enviar um catálogo de produtos para os novos clientes da empresa.
No projeto nos é dados um conjunto de dados histórico, com dados de clientes e a média da receita que aqueles clientes já deram para a empresa.
Temos também o custo de cada catálogo que deve ser levado em consideração antes de tomar a decisão final.
Para que seja possível prever o quanto a empresa pode ter de lucro no envio dos catálogos para os novos clientes é necessária uma análise dos dados anteriores para que seja criada uma fórmula que nos ajude a tomar essa decisão.

Decisões Chaves:

Responda estas perguntas

1. **Que decisões precisam ser feitas??**
   Se o dono da empresa deve ou não investir em catálogos dos produtos para conseguir mais vendas dos clientes novos.

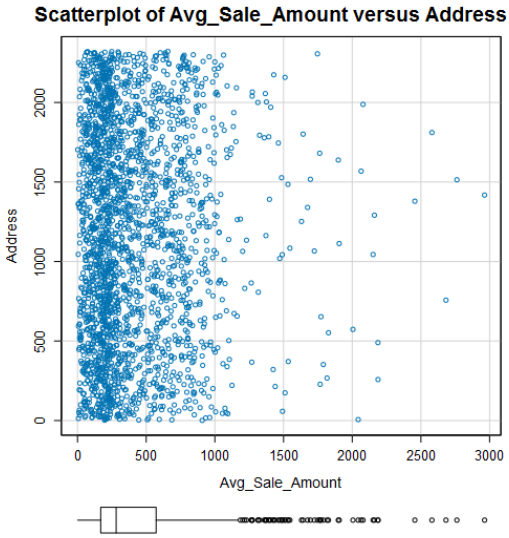
2. **Que dados são necessários para subsidiar essas decisões??**
   Os dados necessários são os presentes no modelo p1-customers.xlsx, mais a frente explicarei detalhadamente quais os dados que utilizei.
   Também é necessário que seja feito um cálculo no valor de cada catálogo que será feito. O lucro deve ser calculado encima da chance do cliente comprar o produto através do catálogo.

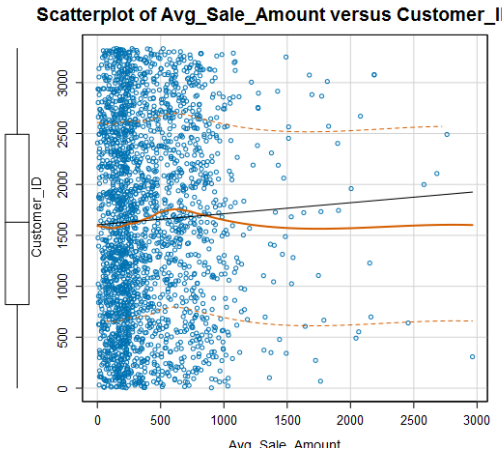
Passo 2: Análise, modelagem e validação

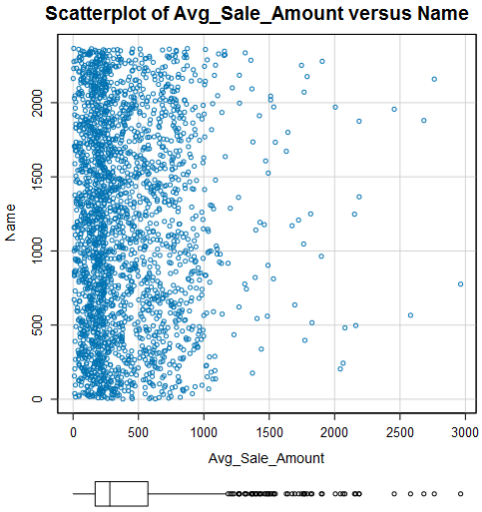
Importante: Use o p1-customers.xlsx para treinar o modelo linear.

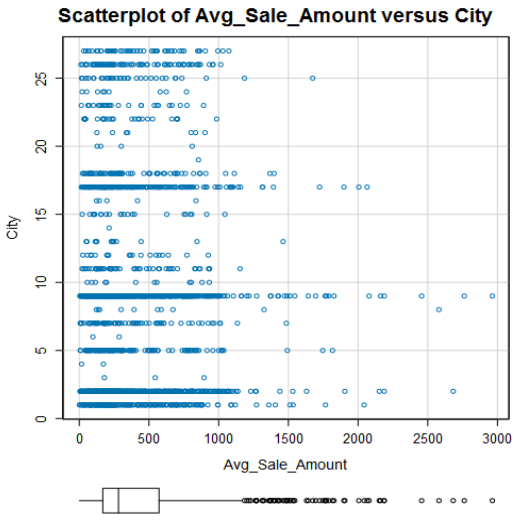
1. **Como e por que você selecionou as variáveis de previsão (veja texto suplementar) em seu modelo? Você deve explicar como as variáveis de previsão contínuas que você escolheu têm uma relação linear com a variável-alvo. Consulte esta lição para ajudar você a explorar seus dados e usar gráficos de dispersão para procurar relações lineares. Você deve incluir gráficos de dispersão em sua resposta.**

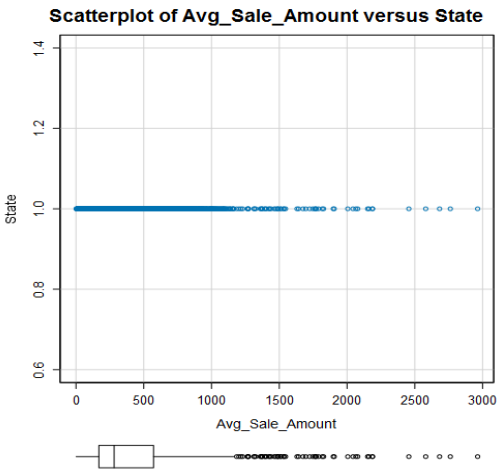
Após analisar os dados presentes no xlsx, verifiquei os seguintes pontos: O nome do cliente seria irrelevante, uma vez que possivelmente ele nem seria repetido no modelo, algo similar ocorre ao estado, o mesmo nunca assume um valor diferente de CO. Os endereços também não se repetirão facilmente. Retirando esses dados do modelo, sobrariam apenas ZIP, Segmento de Clientes, Cidade, Número da Loja, responderam ao ultimo catálogo, AVG de produtos comprados e anos que é cliente.
Abaixo tem imagens que mostram a relação entre a variável alvo e as preditoras:

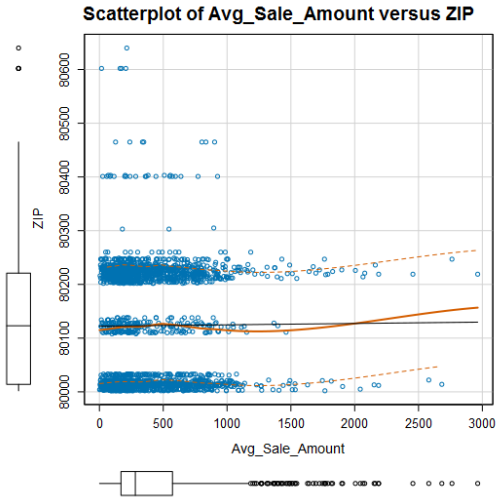


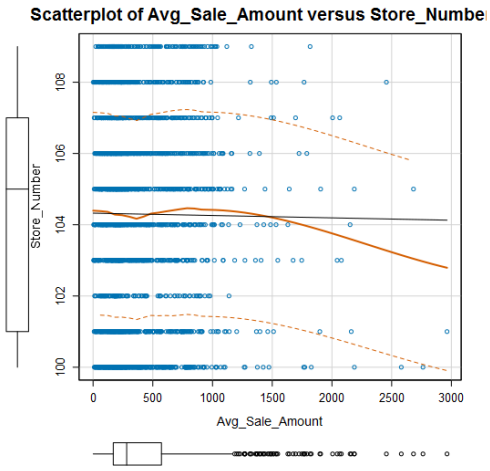


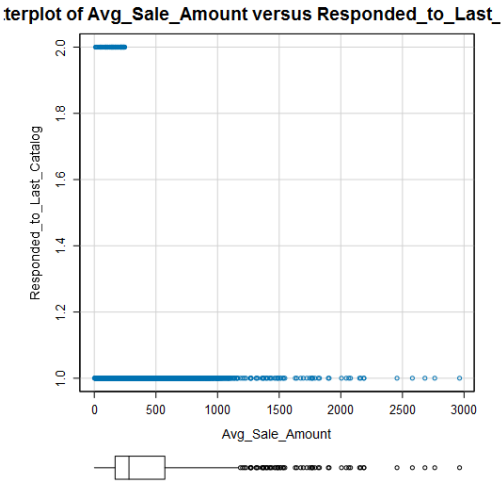


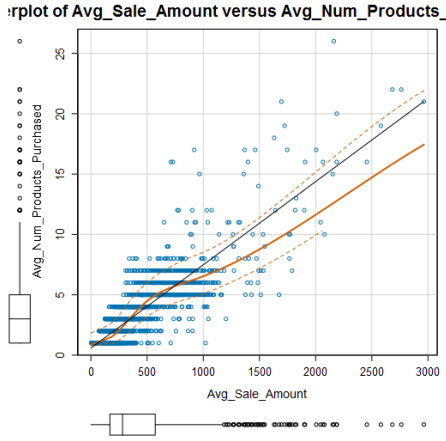


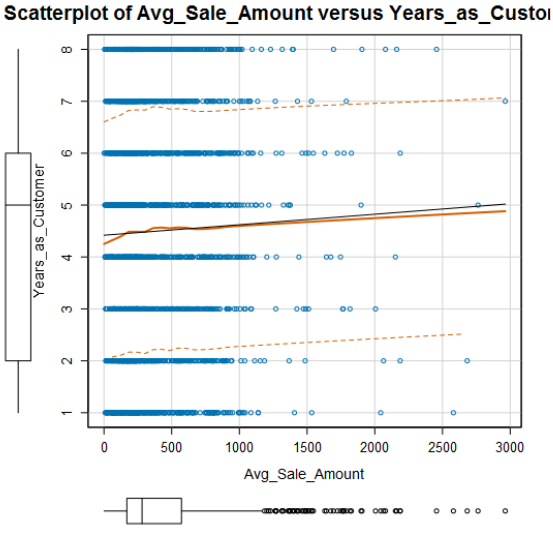



Scatterplot of Avg_Sale_Amount versus ZIP
ZIP
80000
80100
80200
80300
80400
80500
80600
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount


Scatterplot of Avg_Sale_Amount versus Store_Numbe
Store_Number
100
102
104
106
108
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount


terplot of Avg_Sale_Amount versus Responded_to_Last_
Responded_to_Last_Catalog
1.0
1.2
1.4
1.6
1.8
2.0
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount


rplot of Avg_Sale_Amount versus Avg_Num_Products_P
Avg_Num_Products_Purchased
0
5
10
15
20
25
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount


Scatterplot of Avg_Sale_Amount versus Years_as_Custo
Years_as_Customer
1
2
3
4
5
6
7
8
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount

Somente utilizando o gráfico de dispersão percebemos que realmente os gráficos que relacionam: endereço, nome, ID, e estado estão com os resultados dispersos e sem relação, ou possuem uma relação contínua sem variação (no caso do estado).

2. Explique por que você acredita que seu modelo linear é um bom modelo. Você deve justificar o seu raciocínio usando os resultados estatísticos criados pelo seu modelo de regressão. Para cada variável selecionada, por favor justificar por que cada variável é uma boa opção para o seu modelo, usando os valores-p e valores R-quadrado produzidos pelo seu modelo.

As variáveis citadas acima foram retiradas da fórmula de regressão linear, após isso foi feita a primeira fórmula da regressão, e os resultados não foram satisfatórios:
A interseção do modelo ficou com o pvalue de 0.02 e nível de significância de uma estrela.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.189e+04	9505.8914	2.302956	0.02137	*

Uma variável que pode ser considerada significativa é o Segmento de Clientes (3 estrelas) e pvalue abaixo de 0.05.

Customer_SegmentLoyalty Club Only	-1.499e+02	9.0122	-16.633129	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.837e+02	11.9647	23.707798	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.453e+02	9.8382	-24.929693	< 2.2e-16	***

Dentre as variáveis Dummy's criadas para a cidade não existiram resultados que possuíssem um grau de significância maior que uma estrela, veja abaixo algumas imagens.

CityDenver	5.648e+01	27.3871	2.062130	0.03931	*
CityEdgewater	8.488e+01	47.4548	1.788750	0.07378	.
CityEnglewood	3.215e+01	23.9810	1.340677	0.18016	
CityGolden	9.256e+01	57.1187	1.620472	0.10527	
CityGreenwood Village	-2.273e+01	39.9056	-0.569506	0.56907	
CityHenderson	-1.154e+02	157.1146	-0.734334	0.46282	
CityHighlands Ranch	3.564e+00	33.4554	0.106541	0.91516	
CityLafayette	-4.286e+01	62.1797	-0.689359	0.49067	
CityLakewood	5.019e+01	28.8606	1.739207	0.08213	.
CityLittleton	1.478e+00	23.2569	0.063568	0.94932	
CityLone Tree	1.083e+02	138.4959	0.782201	0.43418	
CityLouisville	-2.295e+01	69.3343	-0.331046	0.74064	
CityMorrison	1.043e+02	75.5349	1.381322	0.16731	
CityNorthglenn	4.560e+01	39.9497	1.141544	0.25376	
CityParker	2.762e+01	31.9074	0.865663	0.38676	
CitySuperior	-4.787e+01	46.7553	-1.023859	0.30601	
CityThornton	9.658e+01	39.1145	2.469121	0.01362	*
CityWestminster	1.023e-01	17.5671	0.005825	0.99535	
CityWheat Ridge	2.314e+01	21.8514	1.058828	0.28979	

O ZIP code possui pvalue abaixo de 0.05 porém não é tão significativo para o modelo.

CityWheat Ridge	2.314e+01	21.8514	1.058828	0.28979	
ZIP	-2.672e-01	0.1187	-2.251493	0.02445	*

Store number atingiu um valor maior do que 0.05.

Store_Number	-1.861e+00	1.1513	-1.616848	0.10605	

Por outro lado, outra variável preditora satisfatória é a média de produtos comprados, que possuía um pvalue pequeno e grau de significância muito bom (3 estrelas).
Por último, a variável Anos como Cliente não apresentou um resultado bom o suficiente para ser utilizado no modelo.

Avg_Num_Products_Purchased	6.714e+01	1.5262	43.995599	< 2.2e-16 ***
Years_as_Customer	-2.374e+00	1.2314	-1.927967	0.05398 .

Então no modelo da regressão linear foram utilizados apenas o: Segmento de Cliente e o AVG de produtos comprados.
Após executar novamente o modelo podemos notar que os valores dos pvalues e grau de significância das variáveis preditoras ficaram aceitáveis, o valor do r^2 ajustado também está acima de 0.7, o que faz de nosso modelo, um bom modelo.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.576	28.69	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.98	1.515	44.21	< 2.2e-16 ***

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

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

3. **Qual é a melhor equação de regressão linear com base nos dados disponíveis? Cada coeficiente não deve ter mais de 2 dígitos após o decimal (ex: 1,28)**

   Importante: A equação de regressão deve estar na forma:

Y = 303.46 – 149.36 * Customer_SegmentLoyalty Club Only + 281.84 * Customer_SegmentLoyalty Club and Credit Card - 245.42 * Customer_SegmentStore Mailing List + 66.98 * Avg_Num_Products_Purchased.

Passo 3: Apresentação/Visualização

Use os resultados do modelo para fornecer uma recomendação. (limite de 500 palavras)

No mínimo, responder à estas perguntas:

1. Qual é a sua recomendação? A empresa deve enviar o catálogo para estes 250 clientes?

Sim, a empresa deve enviar o catálogo para os novos clientes. O retorno de investimento ultrapassará os U$ 10.000,00

2. Como você chegou na sua recomendação? (Por favor, explique a sua lógica para os revisores poderem lhe dar feedback sobre o seu processo)
   O primeiro passo foi identificar qual a fórmula de regressão linear seria utilizada no modelo.
   Após isso treinamos os dados no arquivo p1-costumers e utilizamos o score para aplicar a fórmula no modelo p1-mailinglist.
   Será gerado então uma nova coluna na p1-mailinglist (X), então coloquei uma fórmula que multiplica o valor de X pela coluna SCORE_YES, pois é necessário que o cálculo de lucro leve em consideração a chance do cliente comprar através do catálogo.
   Após isso somamos os novos valores de X utilizando a ferramenta Summarize, com isso temos o valor bruto de qual seria o lucro caso não houvessem despesas.
   Para calcular o valor líquido é necessário acrescentarmos o custo de fabricação do catálogo, porém antes disso temos que multiplicar nosso valor bruto por 0.5 pois essa é a margem bruta (custo de fabricação do produto - preço) 50%.
   Após realizado esses cálculos podemos subtrair o valor do custo U$ 6.50(preço de cada catálogo) * 250 (número de e-mails presentes no xlsx).

3. Qual é o lucro esperado do novo catálogo (assumindo que o catálogo é enviado para estes 250 clientes)?
   O valor final do lucro é de U$ 21987.4356865455.