# Projeto 1: Prevendo Demanda de um Catálogo

## Passo 1: Compreensão do Negócio e dos Dados

### Decisões Chaves:

Decisão que devemos prever nesse projeto é se será rentável o envio de catálogo para os 250 novos clientes. Mais especificamente a empresa está interessada em prever quanto poderá lucrar ao enviar esses catálogos.
Para tomar essas decisões, primeiramente devemos compreender os custos associados a confeccionar e enviar esses catálogos e como é definido o cálculo do lucro.
Sabemos que:

- O custo de impressão e distribuição é de US$6,50 por catálogo.
- A margem bruta média (preço - custo) de todos os produtos vendidos através do catálogo é 50%.
- Certifique-se de multiplicar sua receita pela margem bruta antes de subtrair o custo de US$6,50 ao calcular seu lucro.

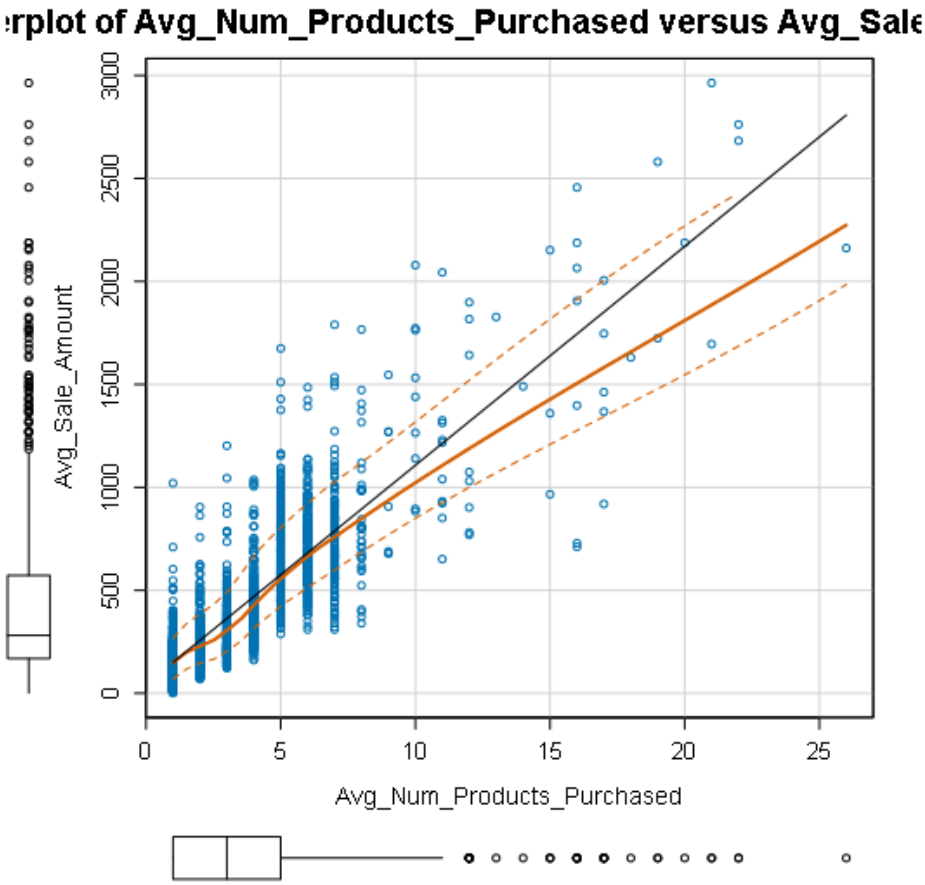
## Passo 2: Análise, modelagem e validação

Dados do cliente contêm as seguintes informações:

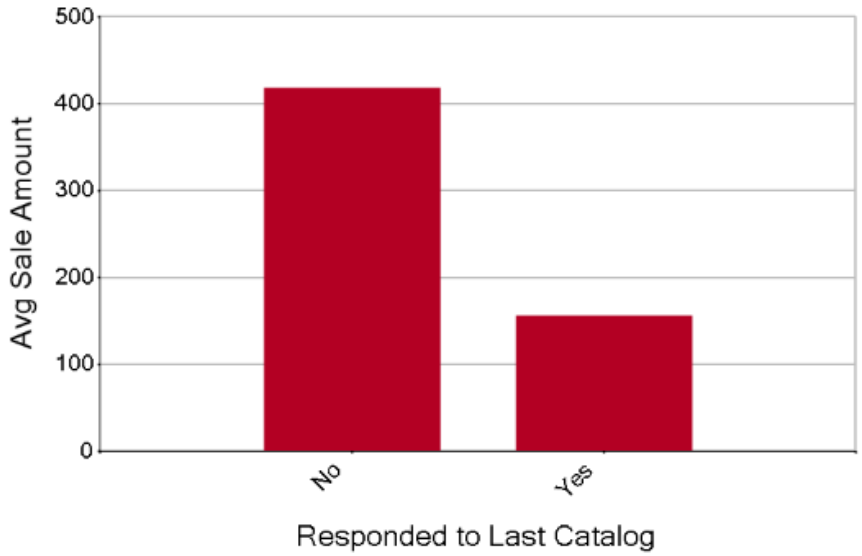
- ID e nome do cliente
- Localização (endereço, cidade, estado e CEP)
- Segmento do cliente
- Número da loja
- Respondeu ao último catálogo
- Média dos produtos
- Quantidade de ano como cliente
- Valor da média de venda (Está será nossa variável target)

Escolher as preditoras para o modelo linear envolve explorar a relação entre cada variável existente em relação a média de valor vendida, utilizando análise bivariada do conjunto de dados. Se uma métrica mostrar um relacionamento linear com a variável target, podemos assumir que funcionará bem como uma entrada da regressão linear.

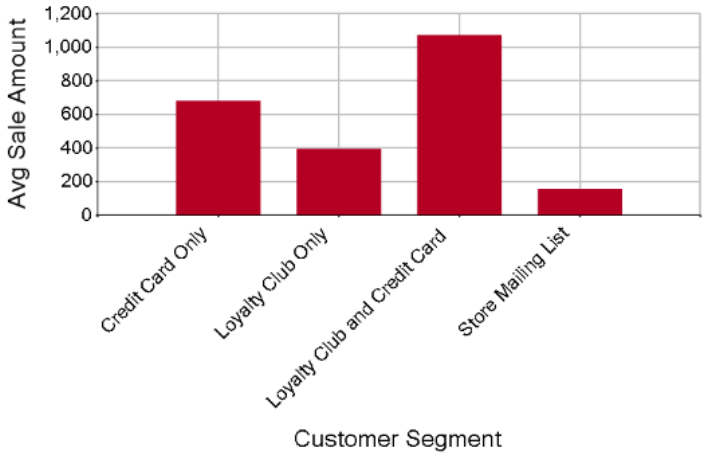
Relação linear mais forte neste conjunto de dados é entre a quantidade média de venda e a quantidade média de produtos comprados.



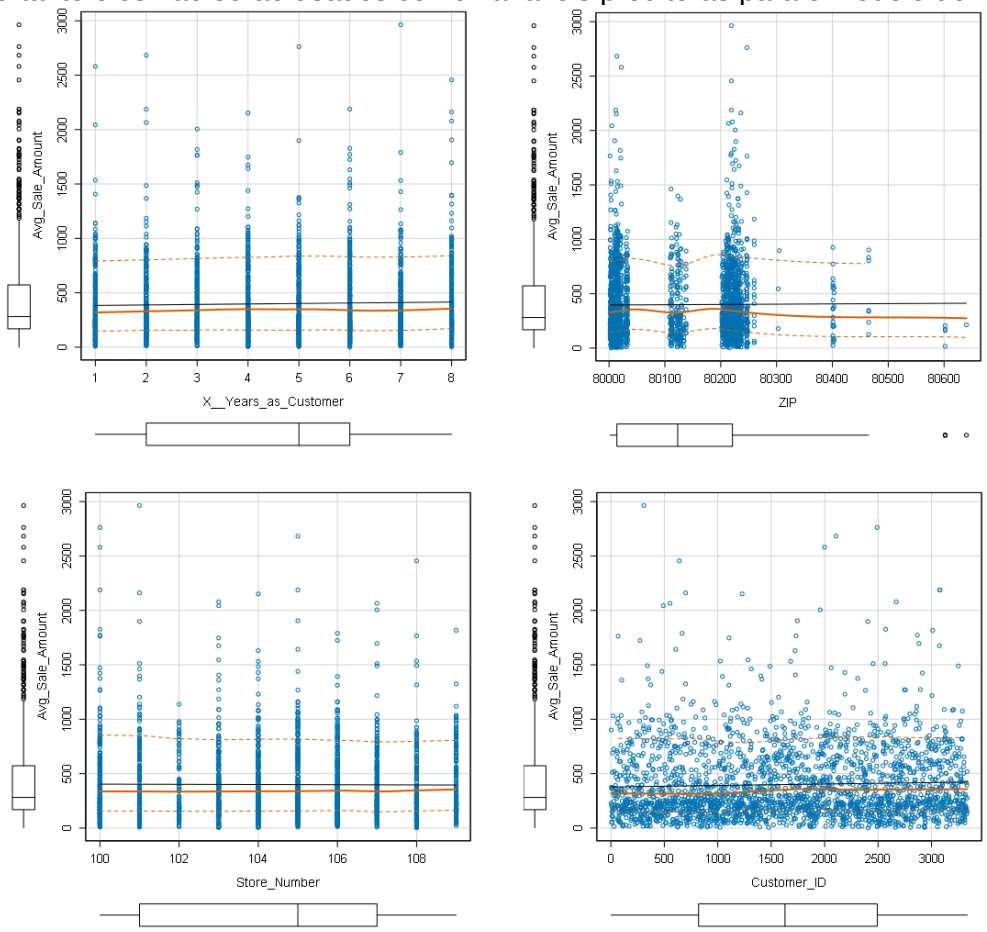

Há também uma relação entre o valor médio da venda e a resposta feita pelo cliente ao último catálogo, onde as pessoas que não responderam ao último catálogo compraram mais que o dobro do pessoal que responderam.



Analisando relação entre valor médio de venda e o segmento do cliente, verificamos também uma forte relação. Clientes que possuem cartão de crédito e são membros do clube de fidelidade tentem a ter maior valor médio de vendas, enquanto os clientes que “Store Mailing List” tendem a ter menor valor.




As métricas restantes, que incluem a quantidade de anos como cliente, CEP, número da loja e número de identificação do cliente não parecem ter muita influência sobre o valor médio de venda. Portanto eles não serão usados como variáveis preditoras para o modelo de regressão.



Primeiro passo para selecionar as preditoras mais eficazes para o modelo linear devemos ajustados todos os dados do cliente em valores numéricos, incluindo os valores binários e variáveis dummy.

Pegaremos a variável dummy “Somente cartão de crédito” do segmento do cliente como caso base e temos os seguintes coeficientes de regressão linear:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.681e+03 | 2.150e+03 | -0.7817 | 0.43445 | |
| Customer.SegmentLoyalty Club Only | -1.504e+02 | 8.974e+00 | -16.7582 | < 2.2e-16 | *** |
| Customer.SegmentLoyalty Club and Credit Card | 2.822e+02 | 1.192e+01 | 23.6760 | < 2.2e-16 | *** |
| Customer.SegmentStore Mailing List | -2.432e+02 | 9.820e+00 | -24.7681 | < 2.2e-16 | *** |
| Customer.ID | -1.622e-03 | 2.939e-03 | -0.5521 | 0.58096 | |
| ZIP | 2.627e-02 | 2.661e-02 | 0.9872 | 0.32365 | |
| Store.Number | -1.012e+00 | 1.006e+00 | -1.0062 | 0.31444 | |
| Responded.to.Last.CatalogYes | -2.891e+01 | 1.128e+01 | -2.5632 | 0.01043 | * |
| Avg.Num.Products.Purchased | 6.683e+01 | 1.517e+00 | 44.0564 | < 2.2e-16 | *** |
| X..Years.as.Customer | -2.315e+00 | 1.222e+00 | -1.8948 | 0.05825 | . |

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

Neste teste determinamos que pegaremos somente os 4 preditores mais significantes conforme print acima, que são:

- Customer.SegmentLoyalty Club Only
- Customer.SegmentLoyalty Club and Credit Card
- Customer.SegmentStore Mailing List
- Avg.Num.Products.Purchased

Construindo o modelo de regressão linear com base nesses quatros recursos resulta numa equação linear como:

Avarage Sale Amount = 303.46 + 66.98 x (Avg Num Products Purchase)
-149. 36 (Se for Customer Segment Loyalty Club Only)
+281.84 (Se for Customer Segment Loyalty and Credit Card)
-245. 42 (Se for Customer Segment Store Mailing List)
+0 (Se for Customer Segment Credit Card Only)

Pontuação do $R^2$ nesse modelo é aproximadamente 0.84, que é um valor alto.

Coefficients:

| | 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
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

# Passo 3: Apresentação/Visualização

Eu recomendaria a empresa enviar o catálogo aos 250 clientes, pois a previsão de lucro supera o lucro esperado pela empresa.

Usando o modelo de regressão linear, multipliquei a pontuação (Score) calculada com a expectativa de venda (Score_Yes) para obter a receita esperada de cada cliente (Expected_Revenue).

| Record # | Name | Customer.Segment | Cust_Id | Address | City | State | ZIP | Store_Number | Avg.Num.Products.Purchased | No_of_years_as_customers | Score_No | Score_Yes | Score | Expected_Revenue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A Giametti | Loyalty Club Only | 2213 | 5326 S Lisbon Way | Centennial | CO | 80015 | 105 | 3 | 0.2 | 0.694964 | 0.305036 | 355.036364 | 108.298804 |
| 2 | Abby Pierson | Loyalty Club and Credit Card | 2785 | 4344 W Roanoke Pl | Denver | CO | 80236 | 101 | 6 | 0.6 | 0.527275 | 0.472725 | 987.159466 | 466.654501 |
| 3 | Adele Hallman | Loyalty Club Only | 2931 | 5219 S Delaware St | Englewood | CO | 80110 | 101 | 7 | 0.9 | 0.421118 | 0.578882 | 622.941184 | 360.609345 |
| 4 | Alejandra Baird | Loyalty Club Only | 2231 | 2301 Lawrence St | Denver | CO | 80205 | 103 | 2 | 0.6 | 0.694862 | 0.305138 | 288.060159 | 87.898046 |
| 5 | Alice Dewitt | Loyalty Club Only | 2530 | 5549 S Hannibal Way | Centennial | CO | 80015 | 104 | 4 | 0.5 | 0.612294 | 0.387706 | 422.012569 | 163.616744 |

Depois disso fiz a somatório da receita esperada dos 250 clientes, calculei a margem bruta média de todos os produtos vendidos que é 50% desse somatório e subtrai o custo dos catálogos (6,50 x 250).

Lucro Esperado = (Somatória da receita esperada x Margem Bruta) – (Custo do catálogo x 250)
= (47,225.87 x 0.5) – (6.50 x 250)
= 23,612.44 – 1,625
= 21,987.44