# Prevendo Demanda de um Catálogo

## Passo 1: Compreensão do Negócio e dos Dados

Estamos prestes a lançar o nosso novo catálogo de vendas, e pretendemos envia-lo a 250 novos clientes. Porém a gerência não quer que enviemos esses catálogos, a menos que o lucro gerado com as vendas realizadas pelo catálogo seja superior a US$10.000.
Com base em um conjunto de dados com 2.375 clientes, iremos criar um modelo de regressão linear que será aplicado a nossa lista de 250 novos clientes. Iremos prever se as compras realizadas por esses clientes através do catálogo nos renderão um lucro igual ou maior que USD$10k.

### Decisões Chaves:

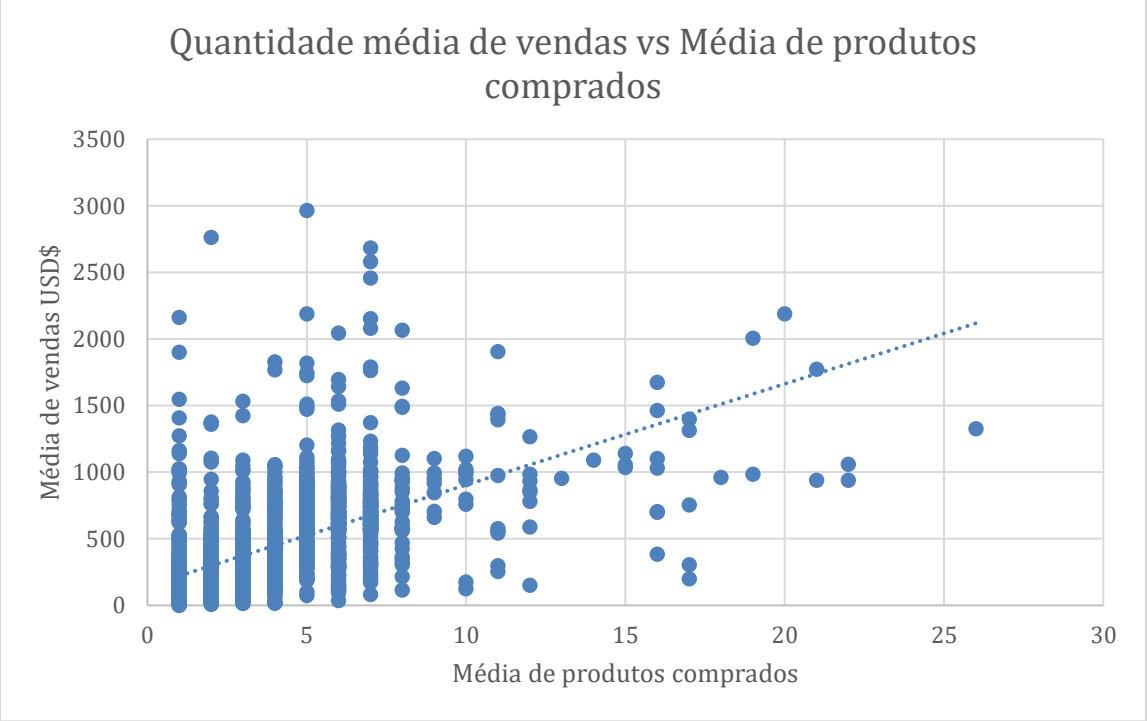
1. Que decisões precisam ser feitas?
   **Resposta:** A empresa está se preparando para enviar o catálogo deste ano nos próximos meses. A empresa tem na sua lista de e-mail, 250 novos clientes para quem eles querem enviar o catálogo. Porém para isso a diretoria precisa saber sobre quanto lucro a empresa terá por enviar um catálogo para esses clientes. A diretoria não quer enviar o catálogo para estes novos clientes a menos que o lucro esperado seja superior a US$10.000. Com base nos resultados gerados em nosso modelo de regressão linear a diretoria poderá decidir se enviará ou não os catálogos aos novos clientes.
2. Que dados são necessários para subsidiar essas decisões?
   **Resposta:** Serão necessários saber algumas características desses novos clientes, em quais segmentos eles se enquadram, qual a probabilidade de comprar ou não através do catálogo. Temos todas essas informações em um conjunto de dados com 2.375 clientes e em nossa lista com os 250 novos clientes. Com base nesses dados iremos criar um modelo de regressão linear. Os coeficientes da regressão linear nos informarão se os valores calculados são confiáveis ou não. A margem de lucro a ser aplicada em todos os produtos vendidos no catálogo é 50% e iremos considerar um custo de impressão e distribuição de USD$6.50 por catálogo, ao todo são 250 catálogos.

## Passo 2: Análise, modelagem e validação

Foi utilizado a ferramenta de regressão linear do software Alteryx.
Usei a variável *Avg Sale Amount* como alvo, e as variáveis *Customer Segment* e *Avg Num Products Purchased* como preditoras. Essas variáveis possuem uma boa relação com a variável alvo, ao configurar o modelo com essas variáveis o mesmo ficou com um R-quadrado ajustado de 0.8366. Sendo um modelo forte para aplicar na regressão linear múltipla.

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.




**Resposta:** Veja nesse gráfico que a variável *Avg Num Products Purchased* tem uma forte relação com *Avg Sale Amount*. A medida que o média de produtos comprados sobe a média de vendas também sobe. Ao aplicar a regressão linear, está variável se mostrou confiável com base no valor-p menor que 0,05. Quando uma variável preditora tem um valor-p abaixo de 0,05, a relação entre ele e a variável alvo é considerada como sendo estatisticamente significativa, ou seja, quanto menor o p-valor, maior a probabilidade de existir uma relação entre o preditor e a variável alvo.

Ao inserir outras variáveis em nosso modelo de regressão percebi que o R-quadrado apresentava um resultado menor que 70%. Um valor R-quadrado próximo de 100% significaria que quase toda a variância na variável alvo é explicada pelo modelo. Um valor R-quadrado próximo de 0% significaria que quase nenhuma variância na variável alvo é explicada pelo modelo. Ao incluir a variável *Customer Segment* percebi que os coeficientes se revelam significativos com um R-quadrado múltiplo de aproximadamente 84% e um valor-p menor que 0,05, indicando uma forte correlação.

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.

   **Resposta:** No modelo aplicado todas as variáveis preditoras tem uma boa relação com a variável alvo, ao configurar o modelo com essas variáveis o mesmo ficou com um R-quadrado ajustado de 0.8366, isso significa que nossas variáveis preditoras tem uma

correlação forte com a variável alvo de aproximadamente 84%. Sendo um modelo forte para aplicar na regressão linear múltipla.

Coeficientes:

| | 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

Erro padrão residual: 137.48 em 2370 graus de liberdade
R quadrada múltipla: 0.8369, R quadrada ajustada: 0.8366
F estatístico: 3040 em 4 e 2370 graus de liberdade (DF), valor p < 2.2e-16

Os valores-p para as variáveis destacadas em vermelho, diz que podemos confiar na estimativa do coeficiente. Com valores-p menor que 0,05 e todas variáveis com uma significância estatística alta, acredito que o modelo seja aceitável para prever o lucro da lista de clientes.

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)
   **Resposta:** *Y = 303.46 + -149.36 * Customer.SegmentLoyalty Club Only + 281.84 * Customer.SegmentLoyalty Club and Credit Card + -245.42 * Customer.SegmentStore Mailing List + 0 * Credit Card + 66,98 * Avg.Num.Products.Purchased*

## Passo 3: Apresentação/Visualização

Com base no modelo de regressão linear criado com um conjunto de 2.375 clientes e, aplicado a lista com 250 clientes. Ao enviar os catálogos aos novos clientes a estimativa de lucro é algo bem maior que os USD$10k. Por isso recomendo que envie os catálogos aos novos clientes.

1. Qual é a sua recomendação? A empresa deve enviar o catálogo para estes 250 clientes?
   **Resposta:** Sim! A empresa deve enviar o catálogo.

2. Como você chegou na sua recomendação? (Por favor, explique a sua lógica para os revisores poderem lhe dar feedback sobre o seu processo)
   **Resposta:** Após construir o modelo de regressão linear e aplicar o modelo a lista de 250 novos clientes e obtive um valor estimado total de USD$ 138,292.13. Para o resultado previsto de vendas multipliquei cada resultado pela variável [score_yes], após esse cálculo o valor total previsto caiu para USD$ 47,224.87. Depois apliquei a renda de 50% sobre as vendas no catálogo e deduzi o custo de impressão e distribuição, obtendo uma previsão de lucro no valor de USD$ 21,987.43.

3. Qual é o lucro esperado do novo catálogo (assumindo que o catálogo é enviado para estes 250 clientes)?
   **Resposta:** USD$ 21,987.43