预测宣传册需求

第 1 步：理解业务和数据

解释下需要作出的关键决策。（限 500 字以内）

关键决策：

请回答以下问题

1. 需要作出什么样的决策？

需要做出的决策：是否向新增的 250 名客户寄送产品目录册。

2. 作出这些决策需要获取哪些数据？

以上决策取决于向新增的 250 名客户寄送产品目录册后，是否能产生超过 1 万美元的利润。

已知：

$$\text{利润}=\sum_{i=1}^{250}(\text{预测收入}_i * \text{毛利率} - \$6.5)$$

预期收入=预测收入*客户购买宣传册中产品的概率（已知）

来自新增的 250 名客户每名客户的预测收入=某线性回归方程（见第 2 步）

因此，需要的数据及其来源如下：

目标	利润	
所需数据	数据名	值
	毛利率	50%
	寄送成本	$6.5/人
	客户购买宣传册中产品的概率	Store_yes!p1-mailinglist
	预测收入 Avg_Sale_Amount	由线性回归方程推算出
线性回归方程所需数据		
	预测变量	参数
		Intercept
	if_Store_Mailing_List	Variable_1
	if_Loyalty_Club_Only	Variable_2
	if_Loyalty_Club_and_Credit_Card	Variable_3
	Avg Num Products Purchased	Variable_4
	Years as Customer	Variable_5
来源	表格 p1-mailinglist：新客户数据	表格 p1-customers：基于老客户数据验证计算得出

第 2 步：分析、建模和验证

描述下你是如何设置线性回归模型的，使用了哪些变量，原因是什么，以及模型的结果。建议提供可视化图表（限 500 字以内）。

在 p1-customers 的老客户数据中，一共包含 12 个变量，筛选如下表：

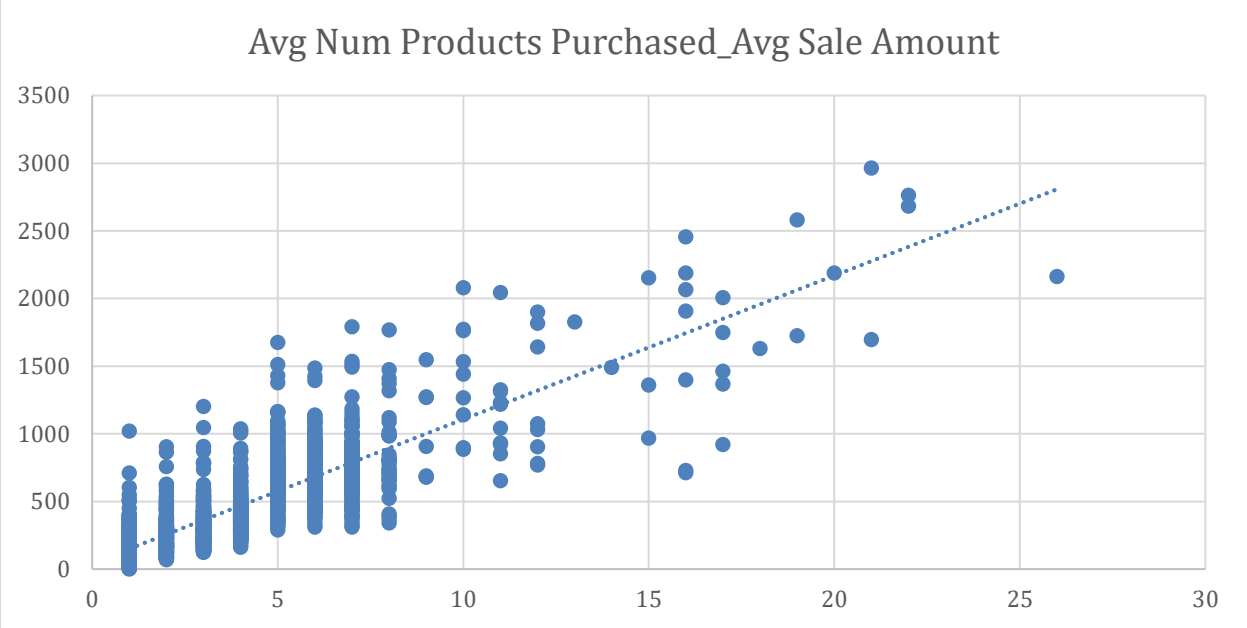
变量	变量类型	筛选原因	是否作为预测变量进行验证
Name	分类变量	因客户而异，分类过多，不宜作为预测变量	否
Customer Segment	分类变量	共4个分类，可以Credit_Card_Only作为基础条件产生3个分类预测变量	是
Customer ID	分类变量	分类过多，不宜作为预测变量	否
Address	分类变量	分类过多，不宜作为预测变量	否
City	分类变量	分类过多，不宜作为预测变量	否
State	分类变量	只有一个分类，无法作为预测变量	否
ZIP	分类变量	分类过多，不宜作为预测变量	否
Avg Sale Amount	数值变量	数值变量，可进行验证	是
Store Number	分类变量	分类过多，不宜作为预测变量	否
Responded to Last Catalog	分类变量	只有一个分类，且在新客户中无相关数据，无法作为预测变量	否
Avg Num Products Purchased	数值变量	数值变量，可进行验证	是
# Years as Customer	数值变量	数值变量，可进行验证	是

待验证的预测变量包括：

1. Avg_Num_Products_Purchased
2. Years_as_Customer
3. 分类变量 if_Store_Mailing_List、if_Loyalty_Club_Only、if_Loyalty_Club_and_Credit_Card（基本条件设为 Only_Credit_Card）

分别验证其与目标变量 Avg_Sale_Amount 之间的线性相关关系：

1. Avg_Num_Products_Purchased：



计算

SUMMARY OUTPUT

回归统计	
Multiple R	0.86
R Square	0.73
Adjusted R Square	0.73
标准误差	176.01
观测值	2375.00

方差分析

	df	SS	MS	F	Significance F
回归分析	1.00	201109435.07	201109435.07	6491.91	0.00
残差	2373.00	73511948.03	30978.49		
总计	2374.00	274621383.09			

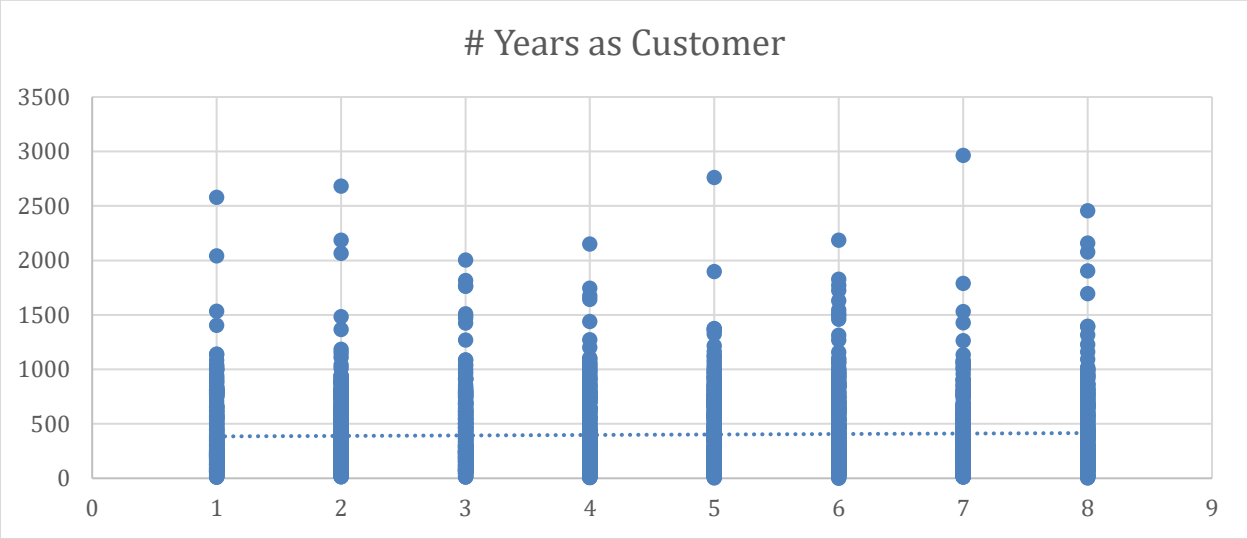
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	44.02	5.70	7.72	0.00	32.83	55.20	32.83	55.20
X Variable 1	106.28	1.32	80.57	0.00	103.69	108.87	103.69	108.87

得 R^2=0.73>0.7，p 值为 0<0.05。

因此：二者显著线性相关，二者的线性相关等式为：

Avg_Sale_Amount=44.02+106.28*Avg_Num_Products_Purchased

2. Years_as_Customer：



计算

SUMMARY OUTPUT

回归统计	
Multiple R	0.03
R Square	0.00
Adjusted R Square	0.00
标准误差	340.04
观测值	2375.00

方差分析

	df	SS	MS	F	Significance F
回归分析	1.00	243578.02	243578.02	2.11	0.15
残差	2373.00	274377805.08	115624.87		
总计	2374.00	274621383.09			

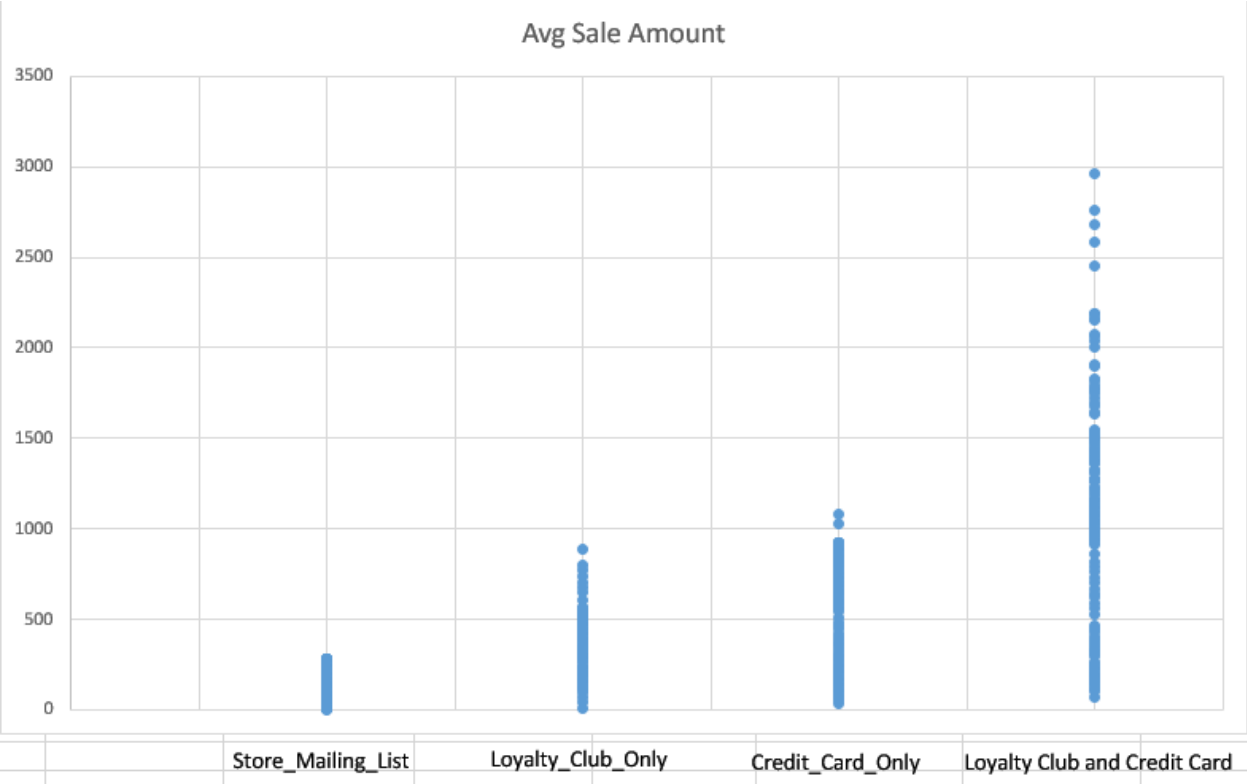
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	380.04	15.28	24.87	0.00	350.07	410.01	350.07	410.01
X Variable 1	4.38	3.02	1.45	0.15	(1.54)	10.31	(1.54)	10.31

得二者的线性相关 R^2=0<0.7，p=0.15>0.05。

可以看出，Years_as_Customer 与目标变量 Avg_Sale_Amount 之间线性相关性不显著，不应选择为预测变量。

3. 分类变量 if_Store_Mailing_List、if_Loyalty_Club_Only、if_Loyalty_Club_and_Credit_Card（基本条件设为 Only_Credit_Card）：

其散点图如下，



计算其回归模型：

SUMMARY OUTPUT

回归统计	
Multiple R	0.84
R Square	0.70
Adjusted R Square	0.70
标准误差	185.67
观测值	2375.00

方差分析

	df	SS	MS	F	Significance F
回归分析	3.00	192884931.52	64294977.17	1865.06	0.00
残差	2371.00	81736451.57	34473.41		
总计	2374.00	274621383.09			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	682.68	8.35	81.72	0.00	666.30	699.06	666.30	699.06
X Variable 1	(525.32)	10.04	(52.30)	0.00	(545.01)	(505.62)	(545.01)	(505.62)
X Variable 2	(286.35)	11.37	(25.18)	0.00	(308.65)	(264.05)	(308.65)	(264.05)
X Variable 3	391.48	15.73	24.89	0.00	360.63	422.33	360.63	422.33

调整的 R^2=0.70，p 值均<0.05 其线性相关性显著。

综上，最终决定选择的预测变量为：

1. Avg_Num_Products_Purchased
2. 分类变量 if_Store_Mailing_List、if_Loyalty_Club_Only、if_Loyalty_Club_and_Credit_Card（基本条件设为 Only_Credit_Card）

回归方程式为：
Avg_Sale_Amount=Intercept+Variable_1(if_Store_Mailing_List*)+Variable_2(if_Loyalty_Club_Only)+Variable_3(if_Loyalty_Club_and_Credit_Card)+0(if_Credit_Card_Only)+Avg_Num_Products_Purchased*Variable_4
代入线性回归模型，计算得：

SUMMARY OUTPUT

回归统计	
Multiple R	0.91
R Square	0.84
Adjusted R Square	0.84
标准误差	137.48
观测值	2375.00

方差分析

	df	SS	MS	F	Significance F
回归分析	4.00	229824514.02	57456128.51	3039.74	0.00
残差	2370.00	44796869.07	18901.63		
总计	2374.00	274621383.09			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.46	10.58	28.69	0.00	282.72	324.20	282.72	324.20
X Variable 1	(245.42)	9.77	(25.13)	0.00	(264.57)	(226.26)	(264.57)	(226.26)
X Variable 2	(149.36)	8.97	(16.65)	0.00	(166.95)	(131.76)	(166.95)	(131.76)
X Variable 3	281.84	11.91	23.66	0.00	258.48	305.19	258.48	305.19
X Variable 4	66.98	1.52	44.21	0.00	64.01	69.95	64.01	69.95

因此，最终的回归方程式为：
Avg_Sale_Amount=303.46+（-245.42）(if_Store_Mailing_List)+（-149.36）(if_Loyalty_Club_Only)+281.84(if_Loyalty_Club_and_Credit_Card)+0(if_Credit_Card_Only)+Avg_Num_Products_Purchased*66.98
同时，整个回归模型的调整的 R^2=0.84，各个变量的 p 均小于 0.05。
因此选择的预测变量与目标变量是显著线性相关的。

第 3 步：演示/可视化:

根据你的模型结果给出建议。（限 500 字以内）

至少回答以下问题：

1. 你的建议是什么？公司应该向这 250 个客户发送宣传册吗？

2. 你是如何得出你的建议的？（请解释你的推理流程，以便审核人员能够根据你的流程向你提供反馈）

3. 新的宣传册带来的利润预计是多少？（假设向这 250 个客户发送了宣传册）

我的建议是公司应该向这 250 个客户发送宣传册。
根据老客户的数据验证得知，每位客户的平均销售可由下述线性回归方程表示，且各预测变量与目标变量线性相关性显著：

Avg_Sale_Amount=303.46+if_Store_Mailing_List*（-245.42）+if_Loyalty_Club_Only*（-149.36）
+if_Loyalty_Club_and_Credit_Card*281.84+Avg_Num_Products_Purchased*66.98
代入已知新客户的类型以及平均产品购买数量（Avg_Num_Products_Purchased），可计算得每位新客户的预测销售额（总额为 138292.1 美元）。
再由每位客户的预测销售额乘以客户购买宣传册中产品的概率，得到每位客户的期望销售额（总额为 47224.87137）。
那么：

$$\text{期望利润}=\sum_{i=1}^{250}(\text{期望收入}_i * 50\% - \$6.5)=21987.44\text{ 美元}>10000\text{ 美元}$$

因此，应当向这 250 个客户发送宣传册，新的宣传册带来的利润预计是 21987.44 美元。