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？
   需要做出的关键决策是：我们是否应该向新增的 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？
   作出这个决策，我们需要知道这 250 个新客户预期能够带来多大的利润。而如果需要知道这个信息，我们就需要知道这 250 个新客户预期能够带来多大的销量，通过销量来计算利润。在我们进行分析的时候，这个销量是不知道的，因此这是一个预测问题，我们需要采用预测分析方法来获得预期销量。根据课程中介绍的方法图，这个预测问题属于有丰富数据的数值预测问题，因此在后面的分析中将采用线性回归模型。而项目已经准备了整洁的数据，因此可以跳过 CRISP-DM 的数据准备步骤，直接开始分析和建模。需要的数据有：

数据名称	数据来源	进一步解释
Avg Sale Amount	p3-customers.xlsx	平均销售额，为用于建模的目标变量
Responded to Last Catalog	p3-customers.xlsx	是否相应过上一次的产品宣传目录，需要进一步建立虚拟变量
Avg Num Products Purchased	p3-customers.xlsx	平均产品购买数量
Customer Segment	p3-customers.xlsx	客户分类，需要进一步建立虚拟变量
# Years as Customer	p3-customers.xlsx	多少年的客户
Customer Segment	p3-mailinglist.xlsx	客户分类，需要进一步建立虚拟变量
Avg Num Products Purchased	p3-mailinglist.xlsx	平均产品购买数量
# Years as Customer	p3-mailinglist.xlsx	多少年的客户
Score_Yes	p3-mailinglist.xlsx	客户决定购买产品的概率
平均毛利率 50%	项目背景提供	平均毛利率信息
产品目录册成本 6.5 美元	项目背景提供	印刷和寄送每本产品目录册的成本

第 2 步：分析、建模和验证

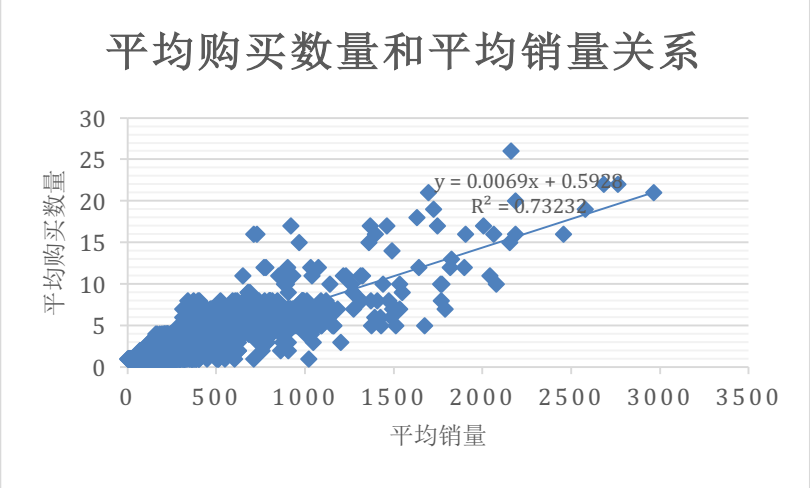
描述下你是如何设置线性回归模型的，使用了哪些变量，原因是什么，以及模型的结果。建议提供可视化图表（限 500 字以内）。

重要事项：使用 ***p1-customers.xlsx 训练你的线性模型。***

至少回答以下问题：

1. 你是如何在你的模型中选择预测变量（请参阅补充文本）的？原因是什么？你必须解释你选择的连续预测变量与目标变量有线性关系。请参阅这节课来探索你的数据，并使用散点图寻找线性关系。你必须在答案中包含散点图。

   首先，我对数据进行了全面的观察，弄清楚给定的数据集中都有哪些字段，每个字段的含义分别是什么，最终我选择了这几个字段作为预测变量：Customer Segment，Responded to Last Catalog，Avg Num Products Purchased，Years as Customer。这几个变量中，只有 Avg Num Products Purchased 和 Years as Customer 是数值变量，其他的两个是分类变量。我们需要通过相关性分析逐一评估和筛选这些变量，首先来看 Avg Num Products Purchased：



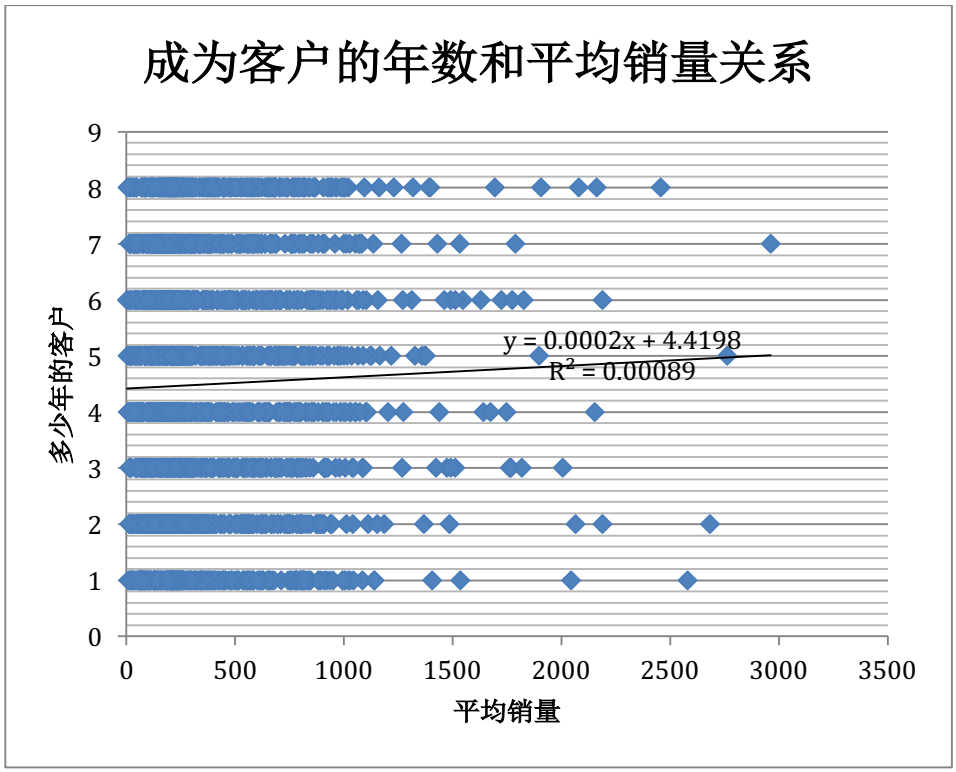
回归统计	
Multiple R	0.855754217
R Square	0.73231528
Adjusted R Square	0.732202476
标准误差	176.0070633
观测值	2375

	Coefficients	*标准误差*	*t Stat*	*P-value*
Intercept	44.01516317	5.704322669	7.71610684	1.75315E-14

X Variable 1	106.2801833	1.319064914	80.57236777	0

从散点图和相关性分析上可以看出 Avg Num Products Purchased 与目标变量有较强相关性，其 R 平方系数为 0.73，系数估计值的 p 值都小于 0.05，具有统计显著性。

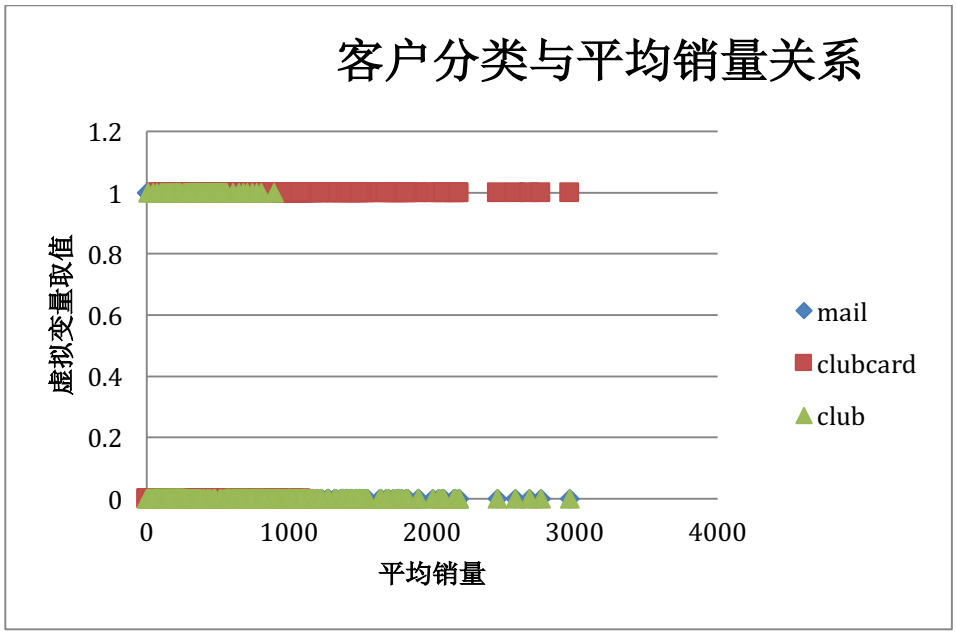
下面再来分析#Years as Customer



回归统计	
Multiple R	0.029781864
R Square	0.000886959
Adjusted R Square	0.000465926
标准误差	340.0365645
观测值	2375

	Coefficients	标准误差	t Stat	P-value
Intercept	380.0388359	15.28292813	24.86688628	1.6908E-121
X Variable 1	4.384997179	3.021175081	1.451421073	0.146794828

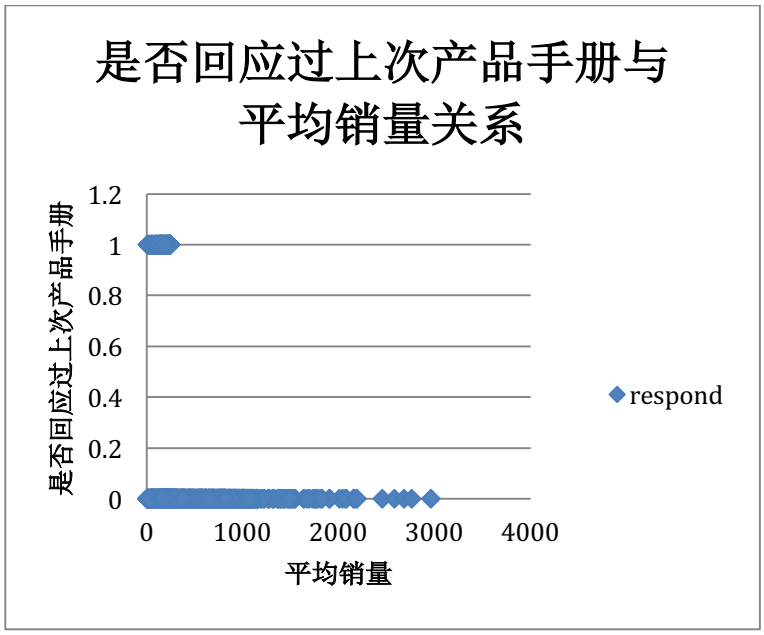
#Years as Customer 与目标变量不具有很显著的相关性，因为 R 平方系数为 0.00088，变量的 p 值 0.15，远大于 0.05，不具备统计显著性，因此排除掉。



回归统计	
Multiple R	0.838073244
R Square	0.702366762
Adjusted R Square	0.70199017
标准误差	185.6701605
观测值	2375

	Coefficients	*标准误差*	*t Stat*	*P-value*
Intercept	682.6789474	8.353695455	81.7217902	0
X Variable 1	-525.3174221	10.0447704	-52.29760376	0
X Variable 2	391.4805372	15.7315673	24.88503082	1.2112E-121
X Variable 3	-286.346374	11.37206197	-25.17981126	3.5029E-124

客户分类和目标变量具有较强相关性，其 R 平方系数为 0.70 且 p 值小于 0.05，具有统计显著性。



回归统计	
Multiple R	0.199358226
R Square	0.039743702
Adjusted R Square	0.039339043
标准误差	333.3587723
观测值	2375

	Coefficients	*标准误差*	*t Stat*	*P-value*
Intercept	418.6566924	7.100780582	58.95924927	0
X Variable 1	-262.2583298	26.46304679	-9.910360355	1.0296E-22

R 方系数只有 0.04，模型不具备可解释性，所以 Responded to Last Catalog 要排除掉。

2. 解释为何你认为你的线性模型是很好的模型。必须使用你的回归模型产生的统计学结果证明你的推理过程。对于你所选择的每个变量，请使用你的模型产生的 p 值和 R 平方值证明每个变量为何与你的模型很好地拟合。

从上面的分析可以看出，我们最后要保留的预测变量为 Avg Num Products Purchased，Customer Segment，而其他的变量要排除掉。利用这里的两个预测变量建立多元线性回归模型，并对其进行分析得到如下结果：

回归统计	
Multiple R	0.914810204
R Square	0.836877709
Adjusted R	0.836602397

Square	
标准误差	137.4832081
观测值	2375

	Coefficients	*标准误差*	*t Stat*	*P-value*
Intercept	303.4634713	10.57571483	28.69436972	1.1227E-155
X Variable 1	66.97620492	1.515040358	44.20753848	0
X Variable 2	-245.4177445	9.767775616	-25.12524388	1.0503E-123
X Variable 3	281.8387649	11.90985741	23.66432739	2.5804E-111
X Variable 4	-149.3557219	8.972754792	-16.64547014	6.34584E-59

从上面两张表可以看出，我得到的多元线性回归模型得到的调整的 R 平方值为 0.84，且对于每个系数估计值，得到的 P 值都很小，小于 0.05，具有统计显著性，意味着得到的结果不大可能是偶然发生的。

3. 根据提供的数据，最佳线性回归方程是什么？每个系数小数点后最多保留两位（例如 1.28）

   根据提供的数据，最佳线性回归方程为：
   Y= 303.46 + 66.98 * Avg_Num_Products_Purchased – 245.42 (If Type: Store Mailiing List) + 281.84 (If Type: Loyalty Club and Credit Card) – 149.36 (If Type: Loyalty Club) + 0 (If Type: Credit Card Only)

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？
   我的建议是公司应该向这 250 个客户发送宣传册。

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）
   首先，我使用上面得到的多元线性回归模型，对 250 个新客户进行了计算，算出每个人预计购买的销售额，然后在此基础上，我进一步计算得到预计的利润，方法是用预计销量乘

以毛利率 50%，然后减去成本 6.5 美元。最后，我将所有的利润加起来得到预计的总利润，超过 1 万美元，因此我建议应该寄送产品宣传册给 250 个新客户。

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）
   新的宣传册带来的利润预计 21987.96 美元。