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。

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）。

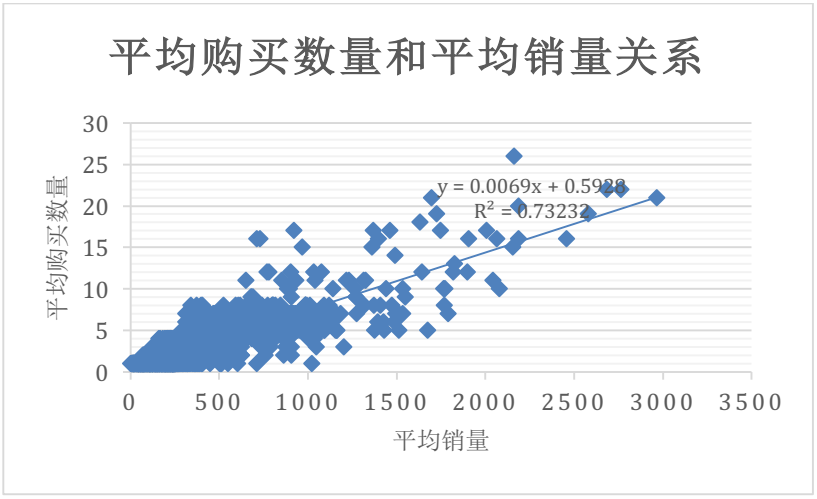
重要事项：使用 ***p1-customers.xlsx 训练你的线性模型。***

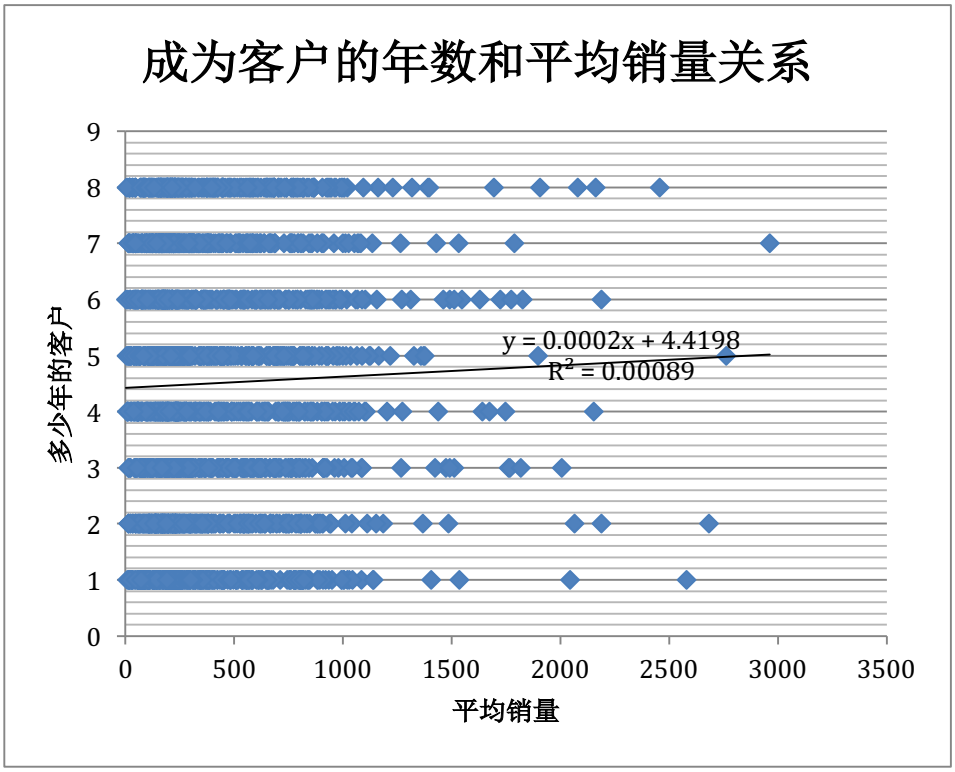
至少回答以下问题：

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，因此我先对数值变量分别画散点图来观察它们与目标变量是否具有线性相关性，得到结果分别是：





平均购买数量和平均销量之间具有明显的线性相关关系，R 方系数为 0.73，然而原先我以为成为客户的年限与销量也有比较明显的相关关系，但是从散点图上看并不是这么回事，R 方系数很小，所以这推翻了我原先的假设，是否是资深客户与平均销量并无关系。至于剩下的两个变量，它们是分类变量，一个表示对客户的分类，有 4 种分类：邮寄产品列表，客户俱乐部，客户俱乐部和信用卡客户，信用卡客户，其中一种分类与邮寄产品列表有关，因此它有一定作用。另一个变量表示该客户有没有回应过上一次邮寄出的产品列表，我认为这也是一个重要的分类变量，所以也把它作为预测变量。为了方便后面的分析和建模，这两个分类变量我已经按要求处理成了 0 和 1 表示的虚拟变量。

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。
3.

回归统计	
Multiple R	0.915045183
R Square	0.837307687
Adjusted R Square	0.836964309
标准误差	137.330867
观测值	2375

	Coefficients	*标准误差*	*t Stat*	*P-value*
Intercept	304.9993073	10.58181267	28.82297361	7.3629E-157
X Variable 1	66.80567645	1.514895347	44.09920234	0
X Variable 2	-242.759476	9.814619383	-24.73447685	2.42E-120
X Variable 3	281.6928352	11.8968034	23.67802726	2.0052E-111
X Variable 4	-150.0315336	8.966880854	-16.73174163	1.74698E-59
X Variable 5	-28.17348048	11.25949084	-2.502198447	0.012409387

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。

4. 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= 305.00 + 66.81 * Avg_Num_Products – 242.76 (If Type: Store Mailiing List) + 281.69 (If Type: Loyalty Club and Credit Card) – 150.03 (If Type: Loyalty Club) + 0 (If Type: Credit Card Only) – 28.17(If Type: Yes) + 0 (If Type: No)

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 22014.60 美元。