预测宣传册需求

请完成每个部分。准备好后，将你的文件另存为 PDF 文档并从课堂上提交。

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？
   需要作出的决策通过预测向 250 名客户寄送产品目录册的盈利，建议管理层是否向新增的他们寄送产品目录册。

2. 作出这些决策需要获取哪些数据？
   向 250 名客户寄送的预计盈利、这些客户购买的概率、预测的购买总价、购买量、成本、毛利率。

第 2 步：分析、建模和验证

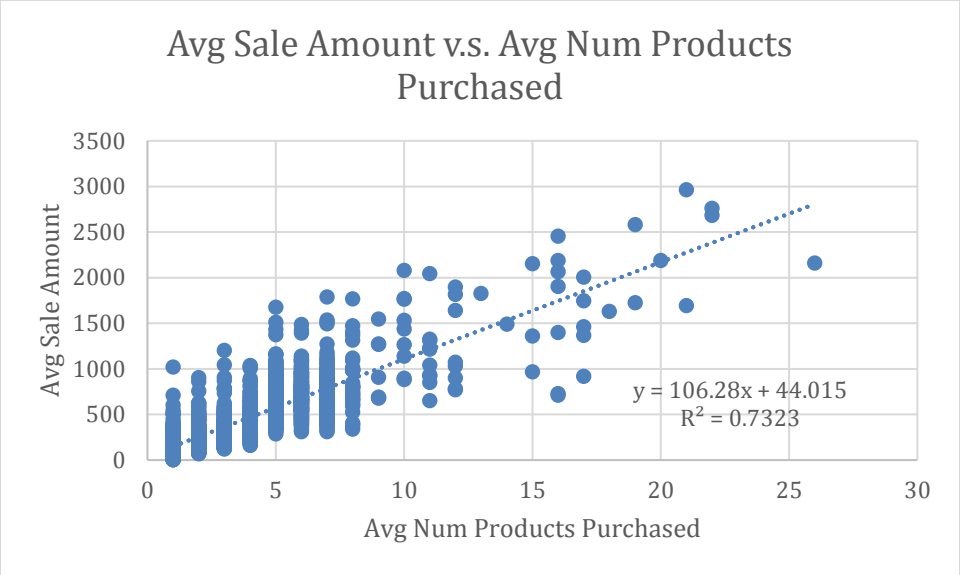
描述下你是如何设置线性回归模型的，使用了哪些变量，原因是什么，以及模型的结果。建议提供可视化图表（限 500 字以内）。

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

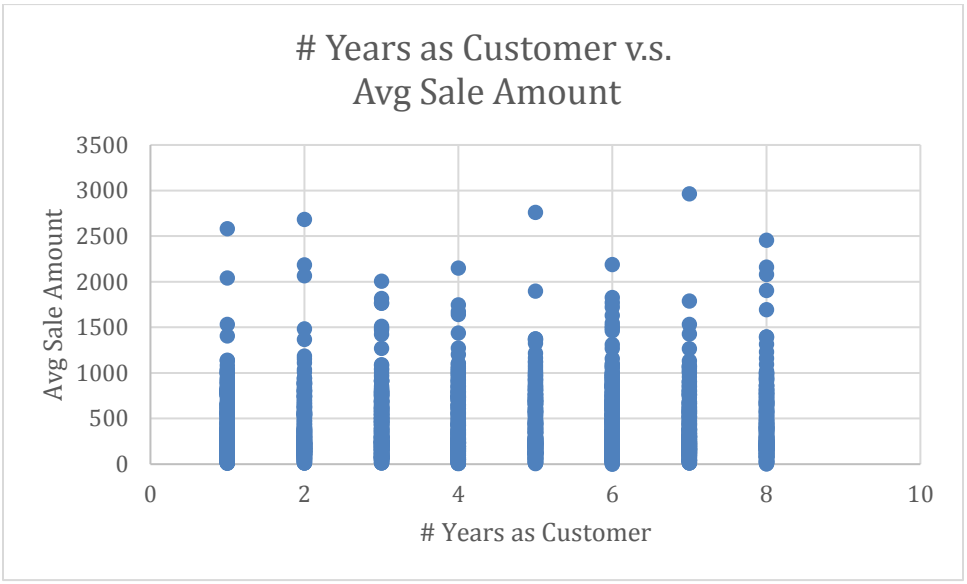
至少回答以下问题：

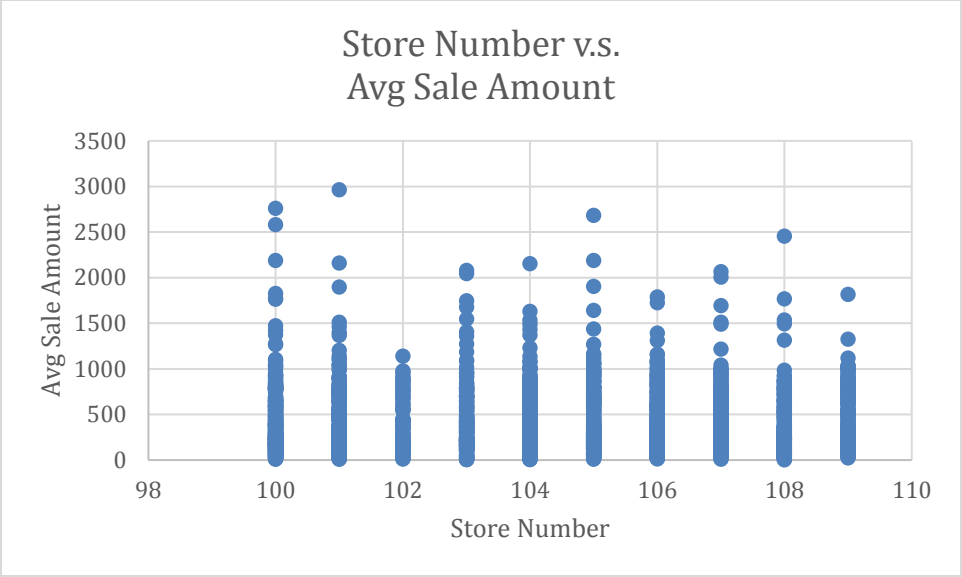
1. 你是如何在你的模型中选择预测变量（请参阅补充文本）的？原因是什么？你必须解释你选择的连续预测变量与目标变量有线性关系。请参阅这节课来探索你的数据，并使用散点图寻找线性关系。你必须在答案中包含散点图。

   我们的目的是要预测新一批客户的购买总价，将这个变量作为目标变量；已有的数据有数值型数据和非数值型数据，首先查看各个数值型数据与目标变量的线性关系，用散点图展示：

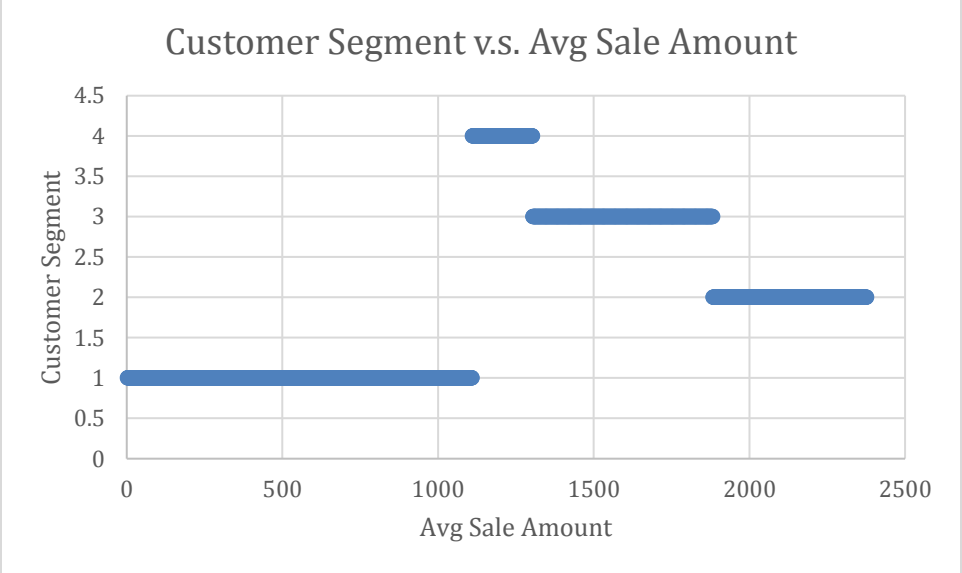


通过散点图发现只有平均购买量与平均购买总价存在线性关系，所以这是构成线性回归模型一个很好的变量。而其他的数值型数据如 number year as customer、store number 不存在线性关系，它们的散点图如下：

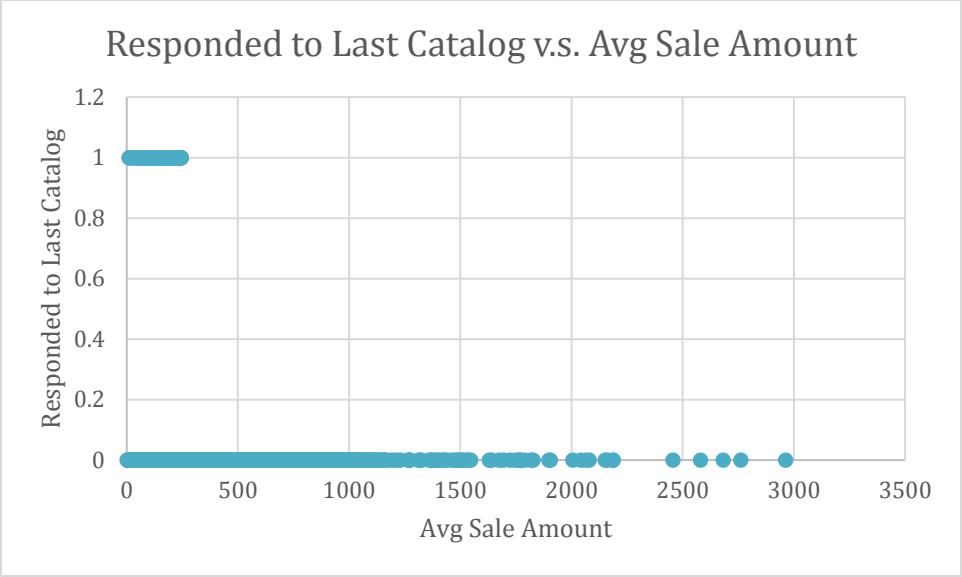




对于非数值型的数据如顾客类别和是否对生产目录有反应的散点图如下：



Store Mailing List	1
Credit Card Only	2
Loyalty Club Only	3
Loyalty Club and Credit Card	4



可以看出顾客类别和是否对生产目录有反应这两个非数值型变量与平均购买总价存在线性关系。因此选择顾客类别、是否对生产目录有反应、平均购买量作为线性模型的预测变量。

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。

用 excel 进行前后 2 次的线性回归分析，目的为了分析是否对生产目录有反应这个预测变量的重要性，发现加入这个变量并不会对调整 R 方改变很大，所以弃掉这个变量。最后计算得到如下模型的结果：

SUMMARY OUTPUT

回归统计	
Multiple	0.91481
R Square	0.836878
Adjusted	0.836602
标准误差	137.4832
观测值	2375

方差分析

	df	SS	MS	F	ignificance F
回归分析	4	2.3E+08	57456129	3039.[illegible]	0
残差	2370	44796869	18901.63		
总计	2374	2.75E+08			

垂直 (值) 轴
绘图区

	Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	585.3022	14.82281	39.4866	1.7E-262	556.2352	614.3692	556.2352	614.3692
X Variabl	66.9762	1.51504	44.20754	0	64.00526	69.94715	64.00526	69.94715
X Variabl	-527.257	13.8613	-38.038	1.5E-247	-554.438	-500.075	-554.438	-500.075
X Variabl	-281.839	11.90986	-23.6643	2.6E-111	-305.194	-258.484	-305.194	-258.484
X Variabl	-431.194	12.69651	-33.9617	2.4E-206	-456.092	-406.297	-456.092	-406.297

R 平方值与调整 R 平方值大约都是 0.837，大于 0.7，反映了预测变量和目标变量的线性关系显著；而且每个预测变量的 p 值都小于 0.05，意味这预测变量与目标变量之间存在不关系的概率较高。

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）

   重要事项：回归方程应该为以下形式

   *Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

   例如： Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

   注意，对于类型 Cash ，我们**必须**包含系数 0。

   Avg Sale Amount = 585.30 + 66.98 * Avg Num Products Purchased – 527.26 * (if customer segment : Store Mailing List) – 281.84 * (if customer segment : Credit Card Only) – 431.19 * (if customer segment : Loyalty Club Only)

注意：如果你使用的是 Alteryx 之外的其他软件，并且决定使用 Customer Segment 作为其中一个预测变量，请将基本条件设为 Only Credit Card。

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？

   我建议公司向这 250 个客户发送宣传册，因为通过线性回归模型预测的销量，再结合客户购买的概率，每本宣传册 6.5 美元的印刷和寄送成本，50%的毛利率，预期净利润是 21,987.44 美元，超过一万美元，所以建议公司发送宣传册给他们。

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）

   通过过去 2375 名客户建立线性回归模型
   Avg Sale Amount = 585.30 + 66.98 * Avg Num Products Purchased – 527.26 * (if customer segment : Store Mailing List) – 281.84 * (if customer segment : Credit Card Only) – 431.19 * (if customer segment : Loyalty Club Only)
   预测这 250 个客户的销量，再乘以客户购买的概率，乘以毛利率，减去成本，计算总的预期例如，计算公式如下：

   Net Profit = Predict Avg Sale Amount * Score_Yes * 50% - 6.5

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）

   利润预计是 21,987.44 美元。

提交之前

请根据此处的审核标准中列出的项目要求检查你的答案。审核人员将根据该审核标准对项目打分。