预测宣传册需求

第 1 步：理解业务和数据

关键决策：

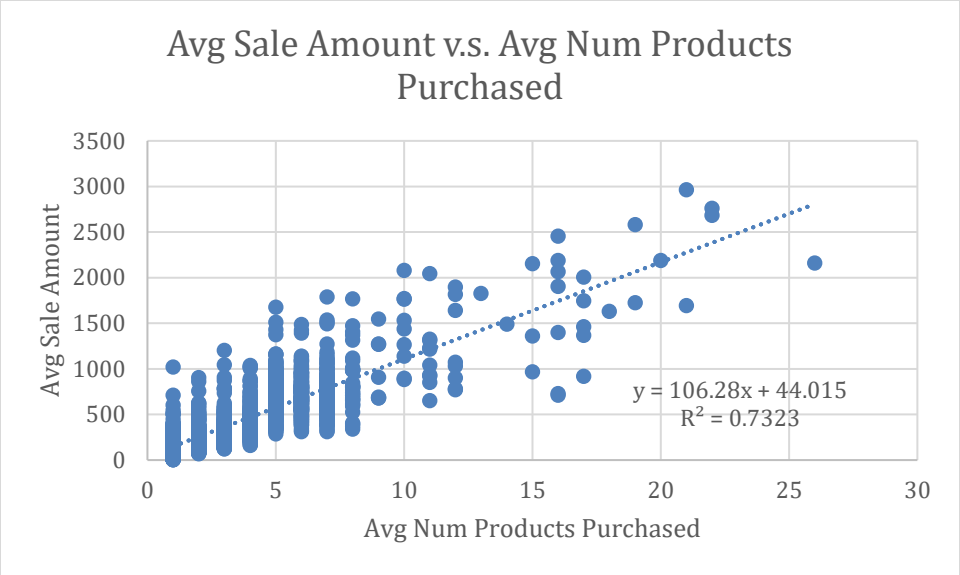
1. 需要作出什么样的决策？
   需要作出的决策是通过预测向 250 名客户寄送产品目录册的盈利，建议管理层是否向新增的他们寄送产品目录册。

2. 作出这些决策需要获取哪些数据？

数据项	数据名称	数据来源	解释
1	Customer Segment	P1_customer.xlsx	在建模过程中建立虚拟变量
2	Avg Num Products Purchased	P1_customer.xlsx	在建模过程中作为数值型数据的预测变量
3	Responded to Last Catalog	P1_customer.xlsx	在建模过程中建立虚拟变量
4	Customer Segment	P1_mailing.xlsx	作为预测变量拟合回归方程来计算预期利润
5	Avg Num Products Purchased	P1_mailing.xlsx	作为预测变量拟合回归方程来计算预期利润
6	Score_Yes	P1_mailing.xlsx	用来计算预期销量
7	印刷/寄送成本	项目信息	用来计算预期净利润
8	平均毛利率	项目信息	用来计算预期净利润
9	寄送给 250 名新客户的的预期净利润	模型预测	用来判断是否值得寄送

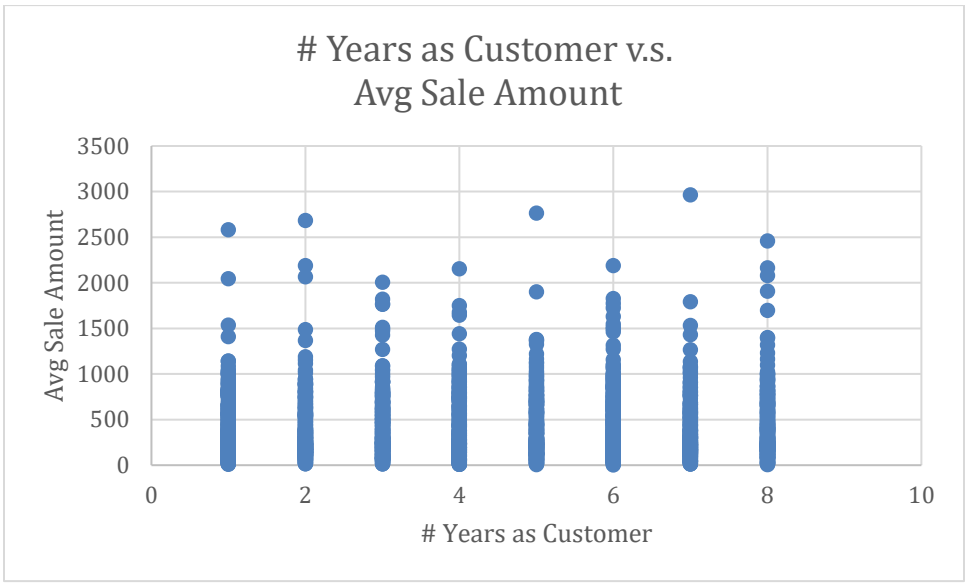
第 2 步：分析、建模和验证

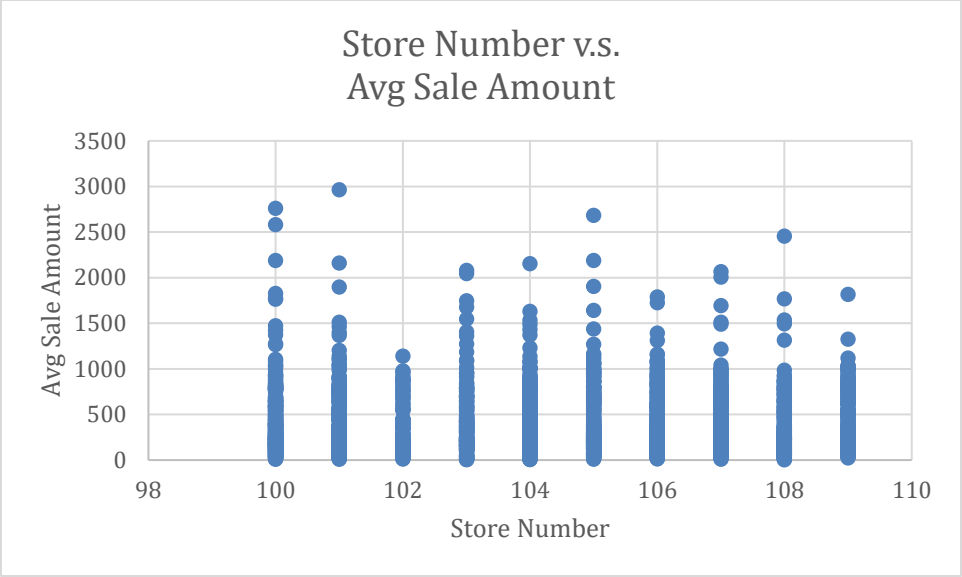
我们的目的是要预测新一批客户的购买总价，将这个变量作为目标变量；已有的数据有数值型数据和非数值型数据，首先查看各个数值型数据与目标变量的线性关系，用散点图展示：



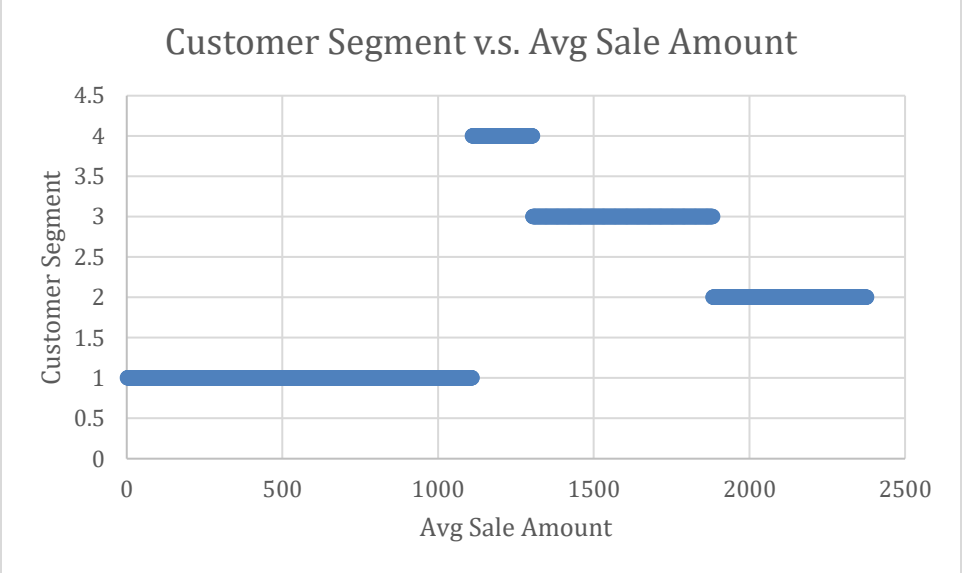
通过散点图发现只有平均购买量与平均购买总价存在线性关系，而且它们之间的相关性 r 为 0.86，R 平方值为 0.73，表明 avg num products purchased 与 目标变量存在较强的线性关系。

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 则不存在线性关系，因为它们与目标变量的相关函数 r 值分别是 0.03、-0.01，它们的散点图如下：

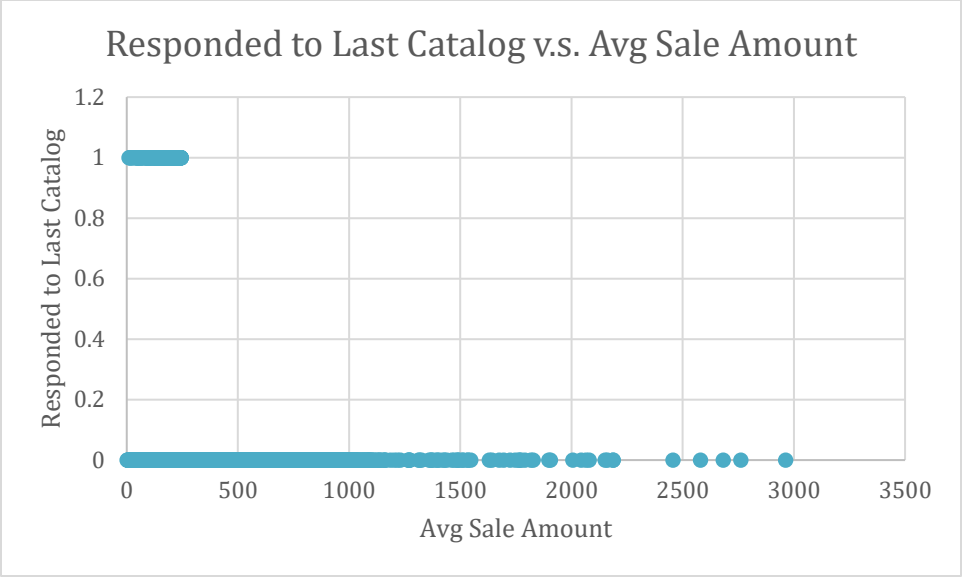




对于非数值型的数据如顾客类别和是否对生产目录有反应的散点图如下：



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。

1. 线性模型：

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	3039.744	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。

2. 根据提供的数据，最佳线性回归方程是：

   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)

第 3 步：演示/可视化:

1. 我的建议：

   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. 建议原因：

   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（假设向这 250 个客户发送了宣传册）：

   21,987.44 美元。