预测宣传册需求

第 1 步：理解业务和数据

关键决策：

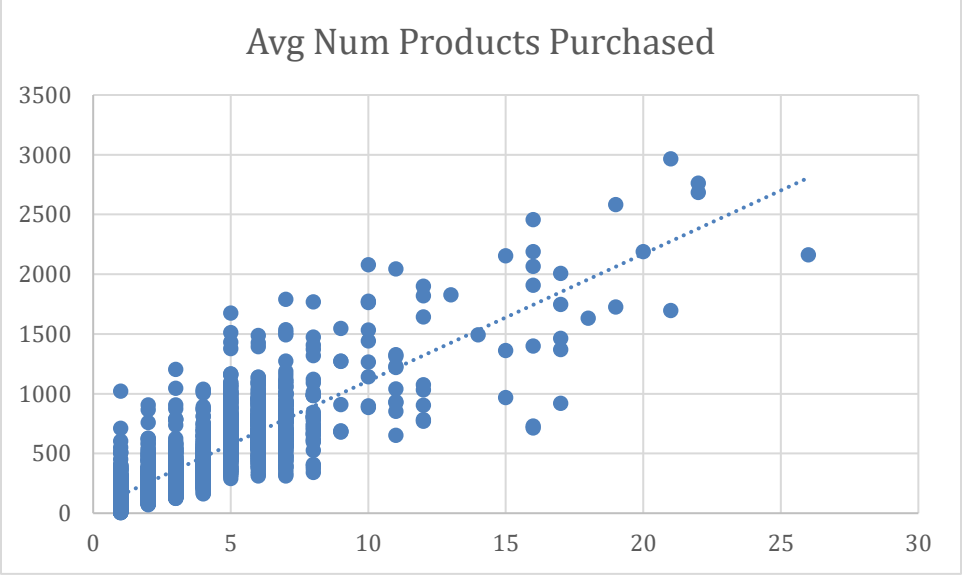
根据 250 个新客户的预期利润，若超过一万美元，管理层将决定向他们寄送产品目录册。
做出这些决策，需要获取的数据是：

- 以往客户的信息，包括客户细分，客户所购买的平均产品数量，客户的平均消费额等。
- 新客户的客户细分，所购买的平均产品数量，以及顾客会对生产目录有所反应且进行购买的概率等。
- 产品目录册的寄送、印刷成本，出售所有产品的毛利率等。

第 2 步：分析、建模和验证

1）寻找连续线性变量

首先，选择连续预测变量，用散点图来寻找线性关系，其中有“Customer ID”，“ Avg Num Products Purchased”，“# Years as Customer”这些可以用来对“Avg Sale Amount”作散点图，其中比较明显具有线性关系的是“ Avg Num Products Purchased”。



SUMMARY OUTPUT

回归统计	
Multiple R	0.855754
R Square	0.732315
Adjusted R Square	0.732202
标准误差	176.0071
观测值	2375

方差分析

	df	SS	MS	F	Significance F
回归分析	1	2.01E+08	2.01E+08	6491.906	0
残差	2373	73511948	30978.49		
总计	2374	2.75E+08			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	44.01516	5.704323	7.716107	1.75E-14	32.82919	55.20114	32.82919	55.20114
Avg Num Products Purchased	106.2802	1.319065	80.57237	0	103.6935	108.8668	103.6935	108.8668

线性变量的 p 值低于 0.05，其与目标变量之间的关系被认为具有统计学意义，是合适的预测变量。

2）寻找分类变量

通过数据透视表，选取有意义的分类变量。

其中可以看出“Customer Segment”为较有意义的分类变量。

SUMMARY OUTPUT

回归统计	
Multiple R	0.838073
R Square	0.702367
Adjusted R Square	0.70199
标准误差	185.6702
观测值	2375

方差分析

	df	SS	MS	F	Significance F
回归分析	3	1.93E+08	64294977	1865.06	0
残差	2371	81736452	34473.41		
总计	2374	2.75E+08			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	682.6789	8.353695	81.72179	0	666.2976	699.0603	666.2976	699.0603
Loyalty Club and Credit Card	391.4805	15.73157	24.88503	1.2E-121	360.6315	422.3296	360.6315	422.3296
Loyalty Club Only	-286.346	11.37206	-25.1798	3.5E-124	-308.647	-264.046	-308.647	-264.046
Store Mailing List	-525.317	10.04477	-52.2976	0	-545.015	-505.62	-545.015	-505.62

分类变量的 p 值均低于 0.05，其与目标变量之间的关系被认为具有统计学意义，是合适的预测变量。

以下为线性回归预测结果

SUMMARY OUTPUT

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

方差分析

	df	SS	MS	F	Significance F
回归分析	4	2.3E+08	57456129	3039.74	0
残差	2370	44796869	18901.63		
总计	2374	2.75E+08			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.4635	10.57571	28.69437	1E-155	282.72486	324.2021	282.7249	324.2021
Loyalty Club and Credit Card	281.8388	11.90986	23.66433	3E-111	258.4839461	305.1936	258.4839	305.1936
Loyalty Club Only	-149.356	8.972755	-16.6455	6.3E-59	-166.950984	-131.76	-166.951	-131.76
Store Mailing List	-245.418	9.767776	-25.1252	1E-123	-264.572015	-226.263	-264.572	-226.263
Avg Num Products Purchased	66.9762	1.51504	44.20754	0	64.00526313	69.94715	64.00526	69.94715

- p 值是系数为零的概率。p 值越低，预测变量和目标变量之间存在关系的概率就越高，结果中预测变量的 p 值均低于 0.05，其与目标变量之间的关系被认为具有统计学意义。

- R 平方可以解读为，模型解释的观察值变差的百分比，R 平方越高，公式在逼近数据方面的表现越好，本方程 R 平方为 0.836878，具有较高的解释力。

回归方程为：
Y = 303.46 + 66.98 * Avg Num Products Purchased + 281.84 (If Type: Credit Card Only) – 149.36 (If Type: Loyalty Club and Credit Card) – 245.42 (If Type: Loyalty Club Only) + 0 (If Type: Store Mailing List)

第 3 步：演示/可视化:

通过回归方程，可算出预测销售总额为：K = 47225.91$
销售利润公式为：P = 0.5 * K - 6.5 $* 250
根据销售利润公式，可得出销售利润为：P = 21987.96$
我的建议是，公司应该向这 250 个客户发送宣传册。新的宣传册带来的利润预计是 21987.96 美元，销售利润大于一万美元。