预测宣传册需求

第 1 步：理解业务和数据

今年新增了 250 名客户，公司希望知道，向这 250 客户寄送产品目录后，这 250 客户的购买商品的利润总和是否能够超过 1 万美元。

数据名称	数据来源	进一步解释
Avg Num Products Purchased	p1-customers	预测变量
# Years as Customer	p1-customers	预测变量
Avg Sale Amount	p1-customers	目标变量
Customer Segment	p1-customers	在建模过程中建立虚拟变量

数据名称	数据来源	进一步解释
Avg Num Products Purchased	p1-mailinglist	代入预测公式计算预测销售额
Customer Segment	p1-mailinglist	代入预测公式计算预测销售额
Score_Yes	p1-mailinglist	预测销售额需要乘以该值

印刷寄送成本：6.5 美元

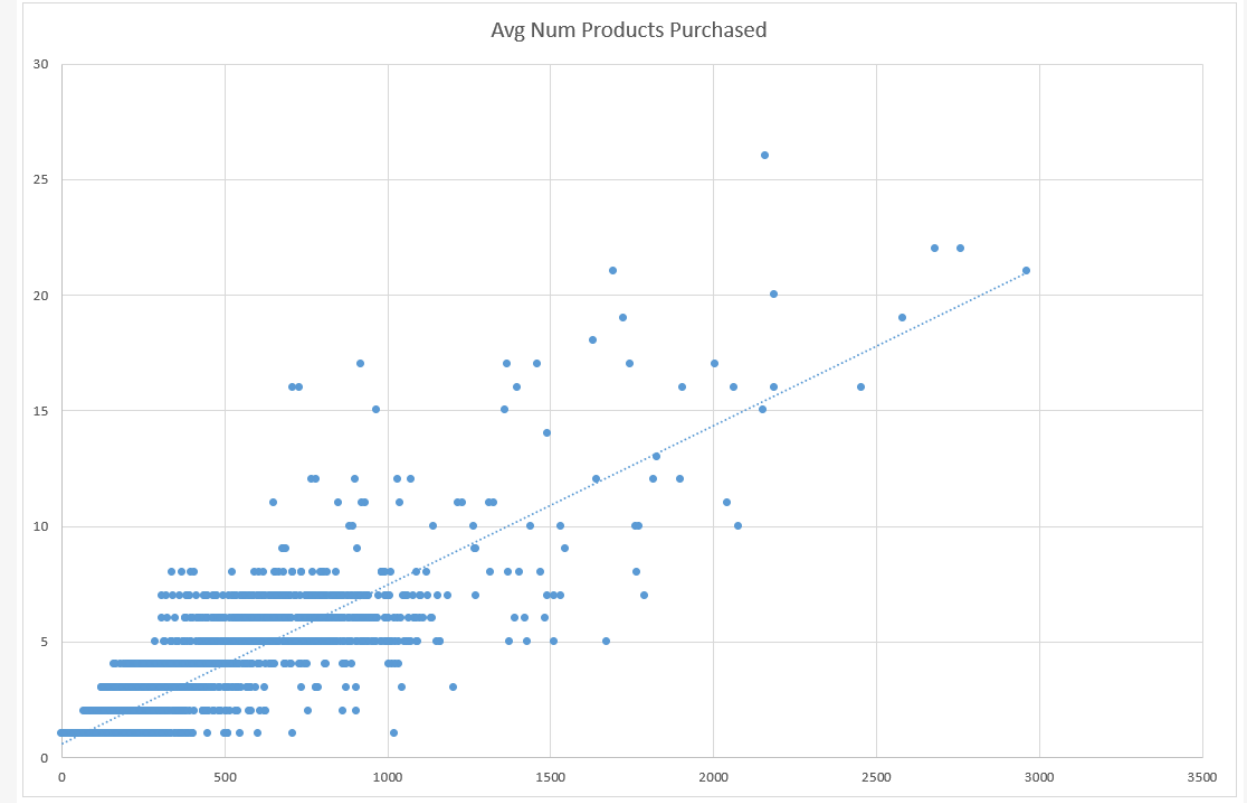
平均毛利率是：50%

关键决策：

1. **公司希望做的决策时是否寄送产品目录。**
2. **做出这个决策需要客户数据的线性回归方程，通过方程预测寄出目录后客户是否能够购买利润总和超过 1 万美元的商品。**

第 2 步：分析、建模和验证

我选择了 Avg Num Products Purchased，# Years as Customer，Customer Segment **作为预测变量。**



SUMMARY OUTPUT

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

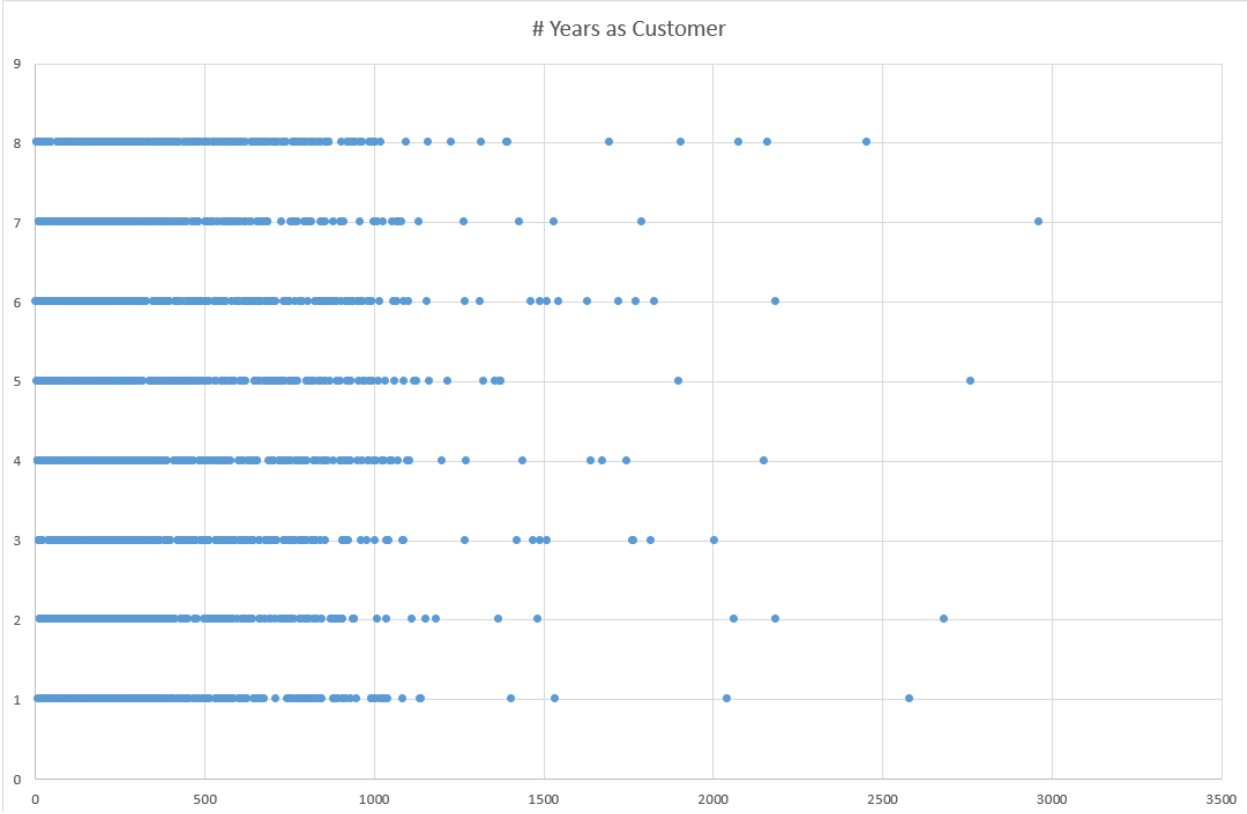
方差分析

	df	SS	MS	F	gnificance 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			

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

1. Avg Num Products Purchased 的 Adjust R square=0.7322，大于 0.7，说明这两个数据之间是强相关的关系,P value 小于 0.05 说明该数据具有统计学意义，由此得出，该数据可以加入目标变量的多元回归方程。

2.接着验证# Years as Customer 与目标变量是否存在关系。



A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
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							
方差分析								
	df	SS	MS	F	gnificance F			
回归分析	1	243578	243578	2.106623132	0.146795			
残差	2373	2.74E+08	115624.9					
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	380.0388359	15.28293	24.86689	1.6908E-121	350.0696	410.0081	350.0696	410.0081
# Years as Customer	4.384997179	3.021175	1.451421	0.146794828	-1.53942	10.30941	-1.53942	10.30941

从散点图看# Years as Customer 跟目标变量之间并没有显著的趋势，下面的数据图显示，Adjust R Square = 0.0004,这个模型并没有什么解释力，下面的 P value = 0.146 远大于 0.05 也说明预测变量与目标变量之间的关系并不具有统计学意义。

3.验证 Customer Segment 与目标变量之间是否存在关系。

B	C	D	E
空	Loyalty Club and Credit Card	Loyalty Club Only	Store Mailing List

因为 Customer Segment 是分类变量，因为有 5 种分类，所以我建立了 4 个虚拟变量，如上图，其中“空”的虚拟变量是因为在原数据集中，Customer Segment 中有些数据是没有分类的，显示为空，所以把空值的也算作一种分类。

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
回归统计								
Multiple R	0.838078008							
R Square	0.702374748							
Adjusted R Square	0.701872427							
标准误差	185.7068356							
观测值	2375							
方差分析								
	df	SS	MS	F	gnificance F			
回归分析	4	1.93E+08	48221781	1398.258501	0			
残差	2370	81734258	34487.03					
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	682.6789474	8.355346	81.70565	0	666.2944	699.0635	666.2944	699.0635
空	-572.128947	185.8947	-3.0777	0.002109889	-936.662	-207.596	-936.662	-207.596
Loyalty Club and Cr	391.4805372	15.73467	24.88012	1.3516E-121	360.6254	422.3357	360.6254	422.3357
Loyalty Club Only	-286.346374	11.37431	-25.1748	3.9179E-124	-308.651	-264.042	-308.651	-264.042
Store Mailing List	-525.275135	10.04815	-52.2758	0	-544.979	-505.571	-544.979	-505.571

对虚拟变量进行回归分析如上图所示，Adjusted R Square =0.7018,这个模型是具有解释力的，P value 都小于 0.05，预测变量与目标变量之间具有统计学意义。

4.经过上面的验证，可以得到 Avg Num Products Purchased，Customer Segment 这两个预测变量与目标变量之间具有统计学意义，并且模型解释力较强，适合将这两个预测变量与目标变量进行多元线性回归。

SUMMARY OUTPUT								
回归统计								
Multiple R	0.914810204							
R Square	0.836877709							
Adjusted R Square	0.836602397							
标准误差	137.4832081							
观测值	2375							
方差分析								
	df	SS	MS	F	gnificance 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						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.4634713	10.57571	28.69437	1.1E-155	282.7249	324.2021	282.7249	324.2021
Avg Num Products Purchased	66.97620492	1.51504	44.20754	0	64.00526	69.94715	64.00526	69.94715
Loyalty Club and Credit Card	281.8387649	11.90986	23.66433	2.6E-111	258.4839	305.1936	258.4839	305.1936
Loyalty Club Only	-149.3557219	8.972755	-16.6455	6.35E-59	-166.951	-131.76	-166.951	-131.76
Store Mailing List	-245.4177445	9.767776	-25.1252	1.1E-123	-264.572	-226.263	-264.572	-226.263

进行多元线性回归分析后可以得到上图数据，可以看到 Adjusted R Square = 0.8366,高于 0.7，预测变量与目标变量之间是强相关关系，在下面数据之中，出了空值外，其它预测变量的 p value 都小于 0.05，具有统计学意义，空值的 p value =0.0594,大于 0.05，不具有统计学意义，可以排除，因此得到多元线性方程为：

y = 303.47

+ 66.98(if type: Avg Num Products Purchased)

+ 281.84(if type: Loyalty Club and Credit Card)

- 149.36(if type: Loyalty Club Only)

- 245.42(if type: Store Mailing List)

+0(if type: `Credit Card Only)`

第 3 步：演示/可视化:

1.我的建议是公司应该向这 250 个客户发送宣传手册。

2.首先已经得到预测销售额公式为 y = 303.47 + 66.98x1 + 281.84x2 + 149.36x3 + 245.41x4

使用 p1-mailinglist 中的 Avg Num Products Purchased，Customer Segment 通过上述预测公式可以得到预测销售额

	K	L
	Score_No	Score_Yes
2	0.694964193	0.305035807

在 p1-mailinglist 中，score_yes 表示购买概率，因此通过预测公式计算的销售额还有乘以购买概率才能的到真正的预测销售额。

因为毛利润=销售额的 50%

所以预测利润 = 预测销售额*0.5 – 6.5（印刷成本和物流成本）
最后得到预测利润为 21987.96 美元。
已经超过原定目标一万美元，所以建议向这 250 名客户寄送手册。