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？
   - 公司是否需要向这 250 个新客户寄送产品目录
   - 如果需要寄送的话，预期盈利是多少
   - 如果预期利润低于 1 万美元，公司不考虑寄送产品目录

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	（进一步）解释
1	Customer Segment	p1-customers.xlsx	在建模过程中建立虚拟变量
2	Avg Num Products Purchased	p1-customers.xlsx	在建模过程中充当预测变量
3	Loyalty Club and Credit Card	p1-customers.xlsx	虚拟变量
4	Loyalty Club Only	p1-customers.xlsx	虚拟变量
5	Store Mailing List	p1-customers.xlsx	虚拟变量
6	# Years as Customer	p1-customers.xlsx	此变量与目标变量无关
7	Credit Card Only	p1-customers.xlsx	基础条件
8	Avg Sale Amount	p1-customers.xlsx	在建模过程中充当目标变量
9	Avg Num Products Purchased	p1-mailinglist.xlsx	利用已建立的模型求预期的销售额
10	Loyalty Club and Credit Card	p1-mailinglist.xlsx	利用已建立的模型求预期的销售额
11	Loyalty Club Only	p1-mailinglist.xlsx	利用已建立的模型求预期的销售额
12	Store Mailing List	p1-mailinglist.xlsx	利用已建立的模型求预期的销售额
13	Credit Card Only	p1-mailinglist.xlsx	利用已建立的模型求预期的销售额
14	Score_Yes	p1-mailinglist.xlsx	预期销售额*Score_Yes表示预期收入

预期利润=预期收入*平均毛利率 – 成本
平均毛利率=50%，寄送成本 = 6.5美元

3. 什么类型的分析能够获取决策所需的信息

   因为我们想要预测 250 人带来的销量，是要预测结果的，又因为是数据充足的连续数值，因此采取线性回归模型。

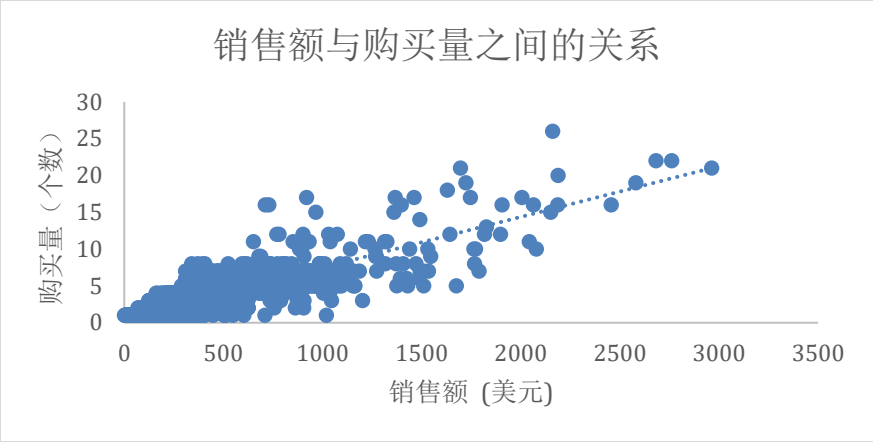
第 2 步：分析、建模和验证

1. 对各个单变量与目标变量绘制散点图

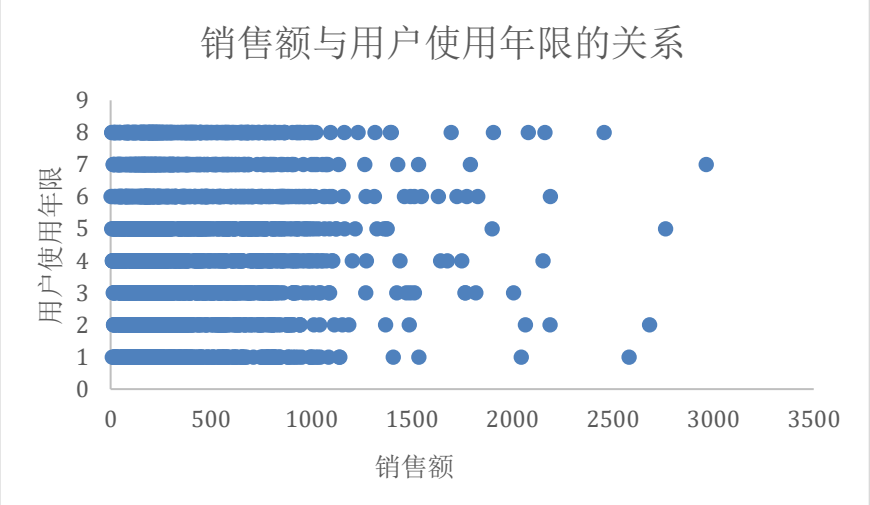
目标变量为 Avg Sale Amount。

预测变量与目标变量的关系：连续变量可以通过散点图观察，虚拟变量与目标变量的关系，可以在回归方程中检验。

通过下面的散点图，可以看出购买量与销售额是存在一定关系的。可以选取购买量作为预测变量



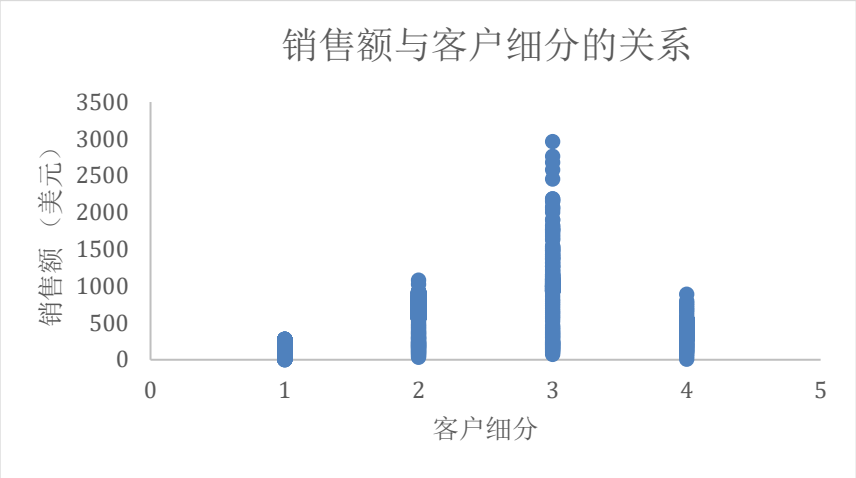
Years as Customer 是连续变量与 Avg Sale Amount 的关系



Customer Segment 的四个变量映射为不同的值

对应分类标签	赋值
Store Mailing List	1
Credit Card Only	2
Loyalty Club and Credit Card	3

Loyalty Club Only	4



2. 对所有涉及到的单变量进行分析和回归

数据项	数据分析对象	数据分析过程	数据分析结论
1	Avg Num Products Purchased	SUMMARY OUTPUT 回归统计 Multiple R 0.855754 R Square 0.73232 Adjusted R Square 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 Coefficient 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	R方大于0.7，P小于0，说明Avg Num Products Purchased对目标变量是有关系的，模型解释力也比较强。Avg Num Products Purchased可以参与多变量线性回归分析。
2	# Years as Customer	# Years as CustomerSUMMARY OUTPUT 回归统计 Multiple R 0.029782 R Square 0.000887 Adjusted R Square 0.000466 标准误差 340.0366 观测值 2375 方差分析 df SS MS F gnificance F 回归分析 1 243578 243578 2.106623 0.146795 残差 2373 2.74E+08 115624.9 总计 2374 2.75E+08 Coefficient 标准误差 t Stat P-value Lower 95% Upper 95%下限 95.0% 上限 95.0% Intercept 380.0388 15.28293 24.86689 1.7E-121 350.0696 410.0081 350.0696 410.0081 # Years as Customer 4.384997 3.021175 1.451421 0.146795 -1.53942 10.30941 -1.53942 10.30941	R方为0.0008，小于0.7，p值大于0.05，预测变量与目标变量显著性无关，此变量不能参与多变量线性回归分析。

3	Loyalty Club and Credit Card	SUMMARY OUTPUT 回归统计 Multiple R 0.591488 R Square 0.349858 Adjusted R 0.349584 标准误差 274.2979 观测值 2375 方差分析 df SS MS F gnificance F 回归分析 1 96078450 96078450 1276.971073 3.8E-224 残差 2373 1.79E+08 75239.33 总计 2374 2.75E+08 Coefficient 标准误差 t Stat P-value Lower 95% Upper 95%下限 95.0%上限 95.0% Intercept 339.7875 5.873468 57.85125 0 328.2698 351.3052 328.2698 351.3052 Loyalty C 734.372 20.55065 35.73473 3.7699E-224 694.0729 774.6711 694.0729 774.6711	R方为0.34，小于0.7，p值小于0.05，预测变量与目标变量有关系，但模型解释力不强。可以参与多变量线性回归分析。
4	Loyalty Club Only	SUMMARY OUTPUT 回归统计 Multiple R 0.005746467 R Square 3.30219E-05 Adjusted R Square -0.000388372 标准误差 340.1818473 观测值 2375 方差分析 df SS MS F gnificance F 回归分析 1 9068.516 9068.516 0.078364 0.779552 残差 2373 2.75E+08 115723.7 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 400.8835802 8.027087 49.94135 0 385.1428 416.6244 385.1428 416.6244 Loyalty Club Only -4.551006776 16.25738 -0.27993 0.779552 -36.4311 27.32913 -36.4311 27.32913	R方大于0.7，p值大于0.05，Loyalty Club Only预测变量与目标变量无显著性关系。
5	Store Mailing List	Store Mailing List SUMMARY OUTPUT 回归统计 Multiple R 0.666655 R Square 0.444429 Adjusted R Square 0.444195 标准误差 253.5642 观测值 2375 方差分析 df SS MS F gnificance F 回归分析 1 1.22E+08 1.22E+08 1898.283 3.3E-305 残差 2373 1.53E+08 64294.82 总计 2374 2.75E+08 Coefficient 标准误差 t Stat P-value Lower 95% Upper 95%下限 95.0%上限 95.0% Intercept 611.7655 7.123602 85.87867 0 597.7964 625.7346 597.7964 625.7346 Store Mailing List -454.404 10.42945 -43.5693 3.3E-305 -474.856 -433.952 -474.856 -433.952	R方为0.44，小于0.7，p值小于0.05，预测变量与目标变量有关系，但模型解释力不强。可以参与多变量线性回归分析。
6	Credit Card Only	SUMMARY OUTPUT 回归统计 Multiple R 0.426357698 R Square 0.181780887 Adjusted R Square 0.181436083 标准误差 307.7181523 观测值 2375 方差分析 df SS MS F gnificance F 回归分析 1 49920919 49920919 527.2011 1.6E-105 残差 2373 2.25E+08 94690.46 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 325.475848 7.095103 45.87331 0 311.5626 339.3891 311.5626 339.3891 Credit Card Only 357.2030994 15.55704 22.96086 1.6E-105 326.6963 387.7099 326.6963 387.7099	R方为0.42，小于0.7，p值小于0.05，预测变量与目标变量有关系，但模型解释力不强。可以参与多变量线性回归分析。

结论：

此模型中，目标变量为 Avg Sale Amount，预测变量建议选取四个，分别是一个连续预测型变量和三个虚拟变量：

- 连续型预测变量是： Avg Num Products Purchased
- 三个虚拟变量：是从 Customer Segment 这个分类型变量分出来的，包含了 Loyalty Club and Credit Card，Loyalty Club Only 和 Store Mailing List。
- 基础条件为 Credit Card Only

3. 筛选完预测变量后，再用所有预测变量进行一次性的多元线性回归求取最终回归方程

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	3039.744	0
残差	2370	44796869	18901.63		
总计	2374	2.75E+08			

	Coefficient	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	1.1E-155	282.72486	324.2021	282.7249	324.2021
Avg Num Products Pur	66.9762	1.51504	44.20754	0	64.00526313	69.94715	64.00526	69.94715
Loyalty Club and Cre	281.8388	11.90986	23.66433	2.6E-111	258.4839461	305.1936	258.4839	305.1936
Loyalty Club Only	-149.356	8.972755	-16.6455	6.35E-59	-166.950984	-131.76	-166.951	-131.76
Store Mailing List	-245.418	9.767776	-25.1252	1.1E-123	-264.572015	-226.263	-264.572	-226.263

从上图可以看出，方程总体拟合优度为 0.8366，且通过了 F 检验，因此回归方程总体显著。从回归系数的检验来看，四个预测变量的 p 值均小于 0.05，表明四个预测变量均对销售额有显著影响。这个线性模型是很好的模型。

最佳回归方程：

Y = 303.46 + 66.98 * 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. 你的建议是什么？公司应该向这 250 个客户发送宣传册吗？

   建议：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？

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Customer Seg	Credit	Loya	Loya	Store M	Avg Nu	预期销售额	Score_No	Score_Yes	预期收入	预期利润	预期总利润						
2	Loyalty Club	0	0	1	0	3	355.04	0.694964	0.3050358	108.30	47.65	21987.96						
3	Loyalty Club	0	1	0	0	6	987.18	0.527275	0.4727245	466.66	226.83							
4	Loyalty Club	0	0	1	0	7	622.96	0.421118	0.5788819	360.62	173.81							
5	Loyalty Club	0	0	1	0	2	288.06	0.694862	0.3051378	87.90	37.45							
6	Loyalty Club	0	0	1	0	4	422.02	0.612294	0.3877059	163.62	75.31							
7	Credit Card	1	0	0	0	7	772.32	0.732722	0.2672783	206.42	96.71		预期销售额==303.46+66.98*G2+281.84*D2-149.36*E2-245.42*F2					
8	Loyalty Club	0	1	0	0	4	853.22	0.778261	0.2217395	189.19	88.10		预期收入=H2*J2					
9	Credit Card	1	0	0	0	6	705.34	0.806553	0.1934471	136.45	61.72		预期利润=K2*50%-6.5					
10	Credit Card	1	0	0	0	6	705.34	0.749342	0.2506576	176.80	81.90		预期总利润==SUM(L2:L251)					
11	Loyalty Club	0	0	1	0	4	422.02	0.735477	0.2645232	111.63	49.32							
12	Store Mailin	0	0	0	1	2	192	0.809459	0.1905414	36.58	11.79							
13	Loyalty Club	0	0	1	0	7	622.96	0.808455	0.1915449	119.32	53.16							

根据回归方程，得到预期销售额，再利用预期收入=预期销售额*是否购买概率，预期利润=预期收入*平均毛利率 – 成本等公式，最终得到 250 个新用户总的预期利润。

将 250 个新用户总的预期总利润与 1 万美元比较，确定是否寄出。新的宣传册带来的利润预计是 21987.96 美元。因为**预期总利润为 21987.96 美元大于 1 万美元**，所以**应该寄出**。