项目：预测邮寄产品目录带来的收入增长

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？

• 计算 250 个新客户的预测利润，来决定是否给他们寄送新的产品目录册。

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	（进一步）数据用途
1	Avg Sale Amount	p1-customers.xlsx	在建立模型中建立被预测变量
2	Avg Num Products Purchased	p1-customers.xlsx	在建模过程中建立预测变量
3	Years as Customer	p1-customers.xlsx	在建模过程中建立预测变量
4	Customer Segment	p1-customers.xlsx	在建模过程中建立预测虚拟变量
5	毛利率	题目	计算真实预测利润
6	寄送成本	题目	计算真实预测利润
7	Avg Num Products Purchased	p1-mailinglist.xlsx	计算新客户的预测收入
8	Years as Customer	p1-mailinglist.xlsx	计算新客户的预测收入
9	Customer Segment	p1-mailinglist.xlsx	计算新客户的预测收入
10	新客户预测利润总和	模型计算	决定是否寄送宣传册

第 2 步：分析、建模和验证

1.模型选择

• 使用线性回归模型，在模型中选择的预测变量是连续型变量 Avg Num Products Purchased 和分类变量 Customer Segment。

理由：

1).连续型变量 Avg Num Products Purchased 与 Avg Sale Amount 存在较显著的线性关系，绘制散点图如图 1，横轴为 Avg Num Products Purchased，纵轴为 Avg Sale Amount，对 Avg Num Products Purchased 和 Avg Sale Amount 进行线性回归结果如图 2，Avg Num Products Purchased 的拟合 p 值为 0<0.05，截距拟合 p 值<0.05，回归结果 R 平方为 0.73，拟合效果较好。

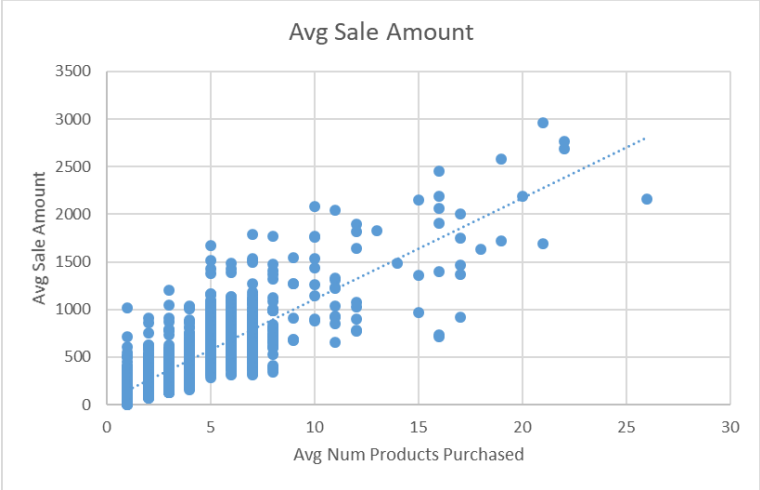

图 1

SUMMARY OUTPUT								
回归统计								
Multiple R	0.855754217							
R Square	0.73231528							
Adjusted R Square	0.732202476							
标准误差	176.0070633							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	1	201109435.1	201109435.1	6491.906448	0			
残差	2373	73511948.03	30978.48632					
总计	2374	274621383.1						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	44.01516317	5.704322669	7.71610684	1.75315E-14	32.82919075	55.20113558	32.82919075	55.20113558
Avg Num Products Purchased	106.2801833	1.319064914	80.57236777	0	103.6935443	108.8668224	103.6935443	108.8668224

图 2

2). 连续型变量 Years as Customer 与 Avg Sale Amount 存在不显著的线性关系，绘制散点图如图 3，横轴为 Years as Customer，纵轴为 Avg Sale Amount，对 Years as Customer 和 Avg Sale Amount 进行线性回归结果如图 4，Years as Customer 的拟合 p 值为 0.14>0.05，截距拟合 p 值<0.05，Years as Customer 作为预测变量不具有统计显著性，回归结果 R 平方为 0.00088，拟合效果不好，因此不选用 Years as Customer 作为预测变量。

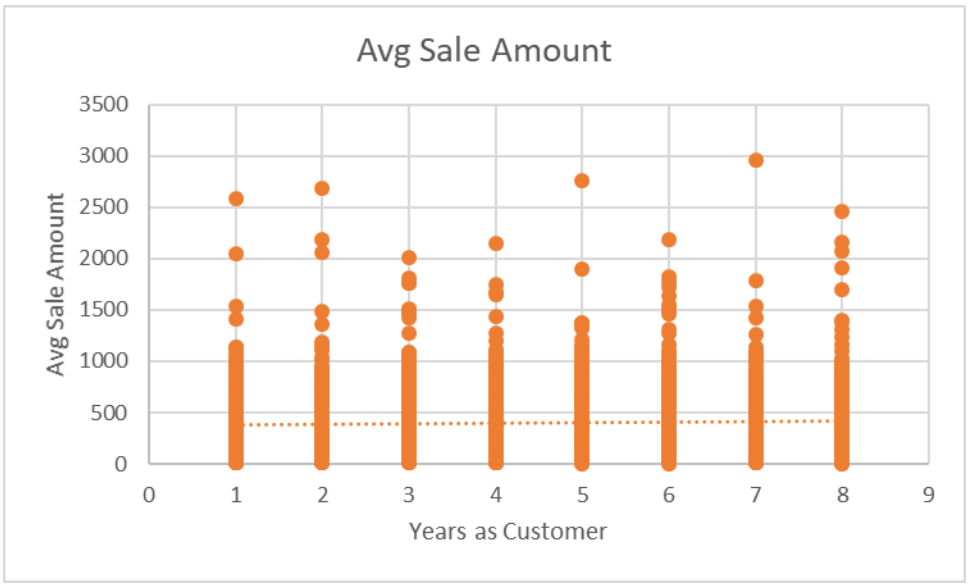

图 3

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	Significance F
回归分析	1	243578.0156	243578.0156	2.106623132	0.146794828
残差	2373	274377805.1	115624.8652		
总计	2374	274621383.1			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	380.0388359	15.28292813	24.86688628	1.6908E-121	350.0695612	410.0081105	350.0695612	410.0081105
# Years as Customer	4.384997179	3.021175081	1.451421073	0.146794828	-1.539418933	10.30941329	-1.539418933	10.30941329

图 4

3). 分类变量 Customer Segment 与 Avg Sale Amount 存在线性相关。对分类变量 Customer Segment 进行处理，将 baseline 设置为 Credit Card Only，将剩余 3 种类型 Loyalty Club and Credit Card，Loyalty Club Only，Store Mailing List 设置为虚拟变量，对 3 个虚拟变量和 Avg Sale Amount 进行线性回归结果如图 5，各虚拟预测变量和截距拟合值 p 值均<0.05，具有统计显著性，模型的 R 平方为 0.7，拟合效果较好。

SUMMARY OUTPUT

回归统计	
Multiple R	0.838073244
R Square	0.702366762
Adjusted R Square	0.70199017
标准误差	185.6701605
观测值	2375

方差分析

	df	SS	MS	F	Significance F
回归分析	3	192884931.5	64294977.17	1865.060055	0
残差	2371	81736451.57	34473.40851		
总计	2374	274621383.1			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	682.6789474	8.353695455	81.7217902	0	666.2976428	699.060252	666.2976428	699.060252
Loyalty Club and Credit Card	391.4805372	15.7315673	24.88503082	1.2112E-121	360.6314839	422.3295904	360.6314839	422.3295904
Loyalty Club Only	-286.346374	11.37206197	-25.17981126	3.5029E-124	-308.6465897	-264.0461582	-308.6465897	-264.0461582
Store Mailing List	-525.3174221	10.0447704	-52.29760376	0	-545.0148655	-505.6199787	-545.0148655	-505.6199787

图 5

2.模型评估

以 Avg Sale Amount 为被预测变量，Avg Num Products Purchased，Loyalty Club and Credit Card，Loyalty Club Only，Store Mailing List 为预测变量进行多元线性回归，回归结果如图 6，在图 6 中，可以看到模型的决定系数 R 平方是 0.837，调整后的 R 平方也是 0.837，说明模型的拟合程度较好，截距以及是各个解释变量的 P 值均小于 0.05，说明具有统计显著性。

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	Significance F
回归分析	4	229824514	57456128.51	3039.744236	0
残差	2370	44796869.07	18901.63252		
总计	2374	274621383.1			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.4634713	10.57571483	28.69436972	1.1227E-155	282.72486	324.2020827	282.72486	324.2020827
Avg Num Products Purchased	66.97620492	1.515040358	44.20753848	0	64.00526313	69.94714671	64.00526313	69.94714671
Loyalty Club and Credit Card	281.8387649	11.90985741	23.66432739	2.5804E-111	258.4839461	305.1935838	258.4839461	305.1935838
Loyalty Club Only	-149.3557219	8.972754792	-16.64547014	6.34584E-59	-166.950984	-131.7604598	-166.950984	-131.7604598
Store Mailing List	-245.4177445	9.767775616	-25.12524388	1.0503E-123	-264.572015	-226.263474	-264.572015	-226.263474

图 6

3.回归方程

- 最佳的线性回归方程为

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:

• 我的建议是公司应该向这 250 个客户发送宣传册.

• 推理流程：根据已经拟合的预测线性回归模型，计算出这 250 名客户的预测收入，乘以顾客购买产品的概率 Score_Yes 后得出真正的预测销售额，用预测销售额乘以毛利率 50%后减去宣传册成本 6.5 得出预测利润，这 250 个客户的预测利润之和超过了一万美元，应该寄送宣传册。

• 假设向这 250 个客户发送了宣传册，新的宣传册带来的利润预计是 21987.44。