预测宣传册需求

第 1 步：理解业务和数据

解释下需要作出的关键决策。（限 500 字以内）

公司今年的邮寄名单中新增了 250 名客户，并希望向这 250 名客户邮寄产品目录册。新客户是否能给公司带来预期利润超过一万美元，若是，将会给 250 名新客户寄送产品目录。

这个决策我将会用到方法图来做分析与解决问题的框架。

查看和分析现有数据后发现，现有数据是数值型和非数值型相混合的数据。需要在数据分析过程中会用到回归模型和分类模型。并根据模型进行建模以及验证模型。（本次分析的数据已是整洁数据，所以不需要进行数据清洗）

最终，根据模型计算结果做出决策。

关键决策：

请回答以下问题

1. 需要作出什么样的决策？

需要做出的决策是，新客户是否能给公司带来预期利润超过一万美元，若是，将会给 250 名新客户寄送产品目录。

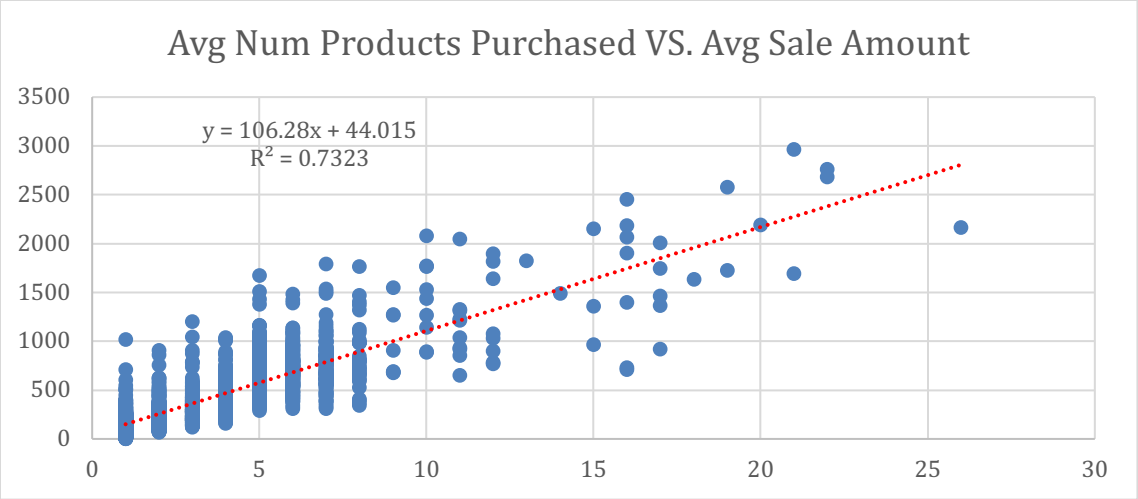
2. 作出这些决策需要获取哪些数据？

数据项	数据名称	数据来源	备注
1	Customer Segment	p1-customers.xlsx p1-mailinglist.xlsx	分类变量需要做虚拟变量
2	Avg Sale Amount	p1-customers.xlsx	回归模型中的应变量
3	Store Number	p1-customers.xlsx p1-mailinglist.xlsx	分类变量需要做虚拟变量
4	Years as Customer	p1-customers.xlsx p1-mailinglist.xlsx	分类变量需要做虚拟变量
5	AvgNumProductsPurchased	p1-customers.xlsx p1-mailinglist.xlsx	数值型自变量
6	Score_Yes	p1-mailinglist.xlsx	可能的利润率
7	6.5 美元	来自公司内部	用来计算利润

第 2 步：分析、建模和验证

描述下你是如何设置线性回归模型的，使用了哪些变量，原因是什么，以及模型的结果。建议提供可视化图表（限 500 字以内）。

分析“**p1-customers.xlsx**”数据的列，发现数值型的自变量“Avg Num Products Purchased”列，对其与应变量”Avg Sale Amount“列做散点图回归模型。（见下图）



SUMMARY OUTPUT

回归统计	
Multiple R	0.85575422
R Square	0.73231528
Adjusted R	0.73220248
标准误差	176.007063
观测值	2375

方差分析

	df	SS	MS	F	gnificance F
回归分析	1	201109435	201109435	6491.90645	0
残差	2373	73511948	30978.4863		
总计	2374	274621383			

	Coefficient:	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	44.0151632	5.70432267	7.71610684	1.75E-14	32.8291907	55.2011356	32.8291907	55.2011356
X Variable	106.280183	1.31906491	80.5723678	0	103.693544	108.866822	103.693544	108.866822

回归模型中的 Multiple R = 0.8558，Adjusted R = 0.7322， P-value.可以看出“Avg Num Products Purchased”与“Avg Sale Amount”是具有相关性的（具有线性关系）。

下面是分类变量，利用 Excel 的 IF 函数，将按照“Years as Customer”的值分别做 7 个列的虚拟变量，然后将这个 7 个虚拟变量与“Avg Sale Amount“做回归模型。得到的结果见下图。

SUMMARY OUTPUT

回归统计	
Multiple R	0.071649
R Square	0.00513358
Adjusted R	0.00219143
标准误差	339.742934
观测值	2375

方差分析

	df	SS	MS	F	gnificance F
回归分析	7	1409790.53	201398.647	1.7448403	0.0943154
残差	2367	273211593	115425.261		
总计	2374	274621383			

	Coefficient:	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	428.743834	19.203406	22.3264474	2.323E-100	391.086594	466.401074	391.086594	466.401074
X Variable	-47.445383	27.5210473	-1.7239672	0.08484434	-101.41324	6.52247502	-101.41324	6.52247502
X Variable	-61.792125	27.0931839	-2.2807259	0.02265318	-114.92096	-8.6632931	-114.92096	-8.6632931
X Variable	-10.156009	27.8167525	-0.3651041	0.71506635	-64.703735	44.3917164	-64.703735	44.3917164
X Variable	-25.952038	27.8423738	-0.9321058	0.35137691	-80.550006	28.6459301	-80.550006	28.6459301
X Variable	-17.382541	27.5930015	-0.629962	0.52878028	-71.491499	36.7264161	-71.491499	36.7264161
X Variable	-4.6360814	27.290087	-0.1698815	0.86511783	-58.151034	48.8788709	-58.151034	48.8788709
X Variable	-65.903081	27.9728693	-2.3559643	0.01855582	-120.75695	-11.049216	-120.75695	-11.049216

由上图可以看出 Multiple R < 0.7， Adjusted R， P-value 可以得出与“Avg Sale Amount”不存在线性关系。

同样的，对“Store Number”分别做 9 个虚拟变量的列，然后做回归模型，得出下图。

SUMMARY OUTPUT

回归统计	
Multiple R	0.05616378
R Square	0.00315437
Adjusted R	-0.0006391
标准误差	340.224479
观测值	2375

方差分析

	df	SS	MS	F	gnificance F
回归分析	9	866257.512	96250.8347	0.831521322	0.58697183
残差	2365	273755126	115752.696		
总计	2374	274621383			

	Coefficient:	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	427.162663	26.1711137	16.3219138	7.83595E-57	375.841958	478.483368	375.841958	478.483368
X Variable	-14.657172	32.2489816	-0.4545003	0.649510485	-77.896379	48.5820349	-77.896379	48.5820349
X Variable	-29.69875	33.2313259	-0.8936974	0.371574747	-94.864302	35.4668025	-94.864302	35.4668025
X Variable	-46.734663	45.2407296	-1.0330219	0.30169929	-135.45027	41.9809406	-135.45027	41.9809406
X Variable	-20.669463	34.6321183	-0.5968293	0.550678485	-88.581923	47.242998	-88.581923	47.242998
X Variable	-40.889811	33.3712583	-1.2253002	0.22058413	-106.32977	24.5501443	-106.32977	24.5501443
X Variable	-7.9218103	32.6258215	-0.242808	0.808175198	-71.899988	56.0563675	-71.899988	56.0563675
X Variable	-45.140684	33.0748745	-1.3648029	0.172444842	-109.99944	19.7180721	-109.99944	19.7180721
X Variable	-13.159742	34.5992379	-0.3803478	0.703721449	-81.007726	54.6882409	-81.007726	54.6882409
X Variable	-67.830329	35.1586477	-1.9292645	0.053817547	-136.7753	1.11463832	-136.7753	1.11463832

同样可以看出 Multiple R < 0.7， Adjusted R， P-value 可以得出与“Avg Sale Amount”不存在线性关系。

将分类变量“Customer Segment”，做 3 列虚拟变量后，做回归模型可得下图，

SUMMARY OUTPUT

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

方差分析

	df	SS	MS	F	ignificance 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
X Variable 1	-525.317422	10.0447704	-52.2976038	0	-545.014866	-505.619979	-545.014866	-505.619979
X Variable 2	-286.346374	11.37206197	-25.1798113	3.5029E-124	-308.64659	-264.046158	-308.64659	-264.046158
X Variable 3	391.4805372	15.7315673	24.88503082	1.2112E-121	360.6314839	422.3295904	360.6314839	422.3295904

由上图的 Multiple R = 0.838， Adjusted R = 0.712， P-value < 0.05,可以得出“Customer Segment”与“Avg Sale Amount“存在线性关系。

下面是将“Customer Segment“的虚拟变量，“Avg Num Products Purchased”和“Avg Sale Amount“做线性回归求得回归方程。

SUMMARY OUTPUT

回归统计	
Multiple R	0.9148102
R Square	0.83687771
Adjusted R	0.8366024
标准误差	137.483208
观测值	2375

方差分析

	df	SS	MS	F	gnificance F
回归分析	4	229824514	57456128.5	3039.74424	0
残差	2370	44796869.1	18901.6325		
总计	2374	274621383			

	Coefficient:	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.463471	10.5757148	28.6943697	1.123E-155	282.72486	324.202083	282.72486	324.202083
X Variable	66.9762049	1.51504036	44.2075385	0	64.0052631	69.9471467	64.0052631	69.9471467
X Variable	-245.41774	9.76777562	-25.125244	1.05E-123	-264.57201	-226.26347	-264.57201	-226.26347
X Variable	-149.35572	8.97275479	-16.64547	6.3458E-59	-166.95098	-131.76046	-166.95098	-131.76046
X Variable	281.838765	11.9098574	23.6643274	2.58E-111	258.483946	305.193584	258.483946	305.193584

由上图的 Multiple R = 0.915， Adjusted R = 0.837， P-value < 0.05,可以得出“Customer Segment“的虚拟变量，“Avg Num Products Purchased”和“Avg Sale Amount“是存在线性关系的。

由上图可以得到线性回归方程为（基本条件为 Only Credit Card），

y = 303.46 + (66.98 * Avg Num Products Purchased) - 245.42(if type: Store Mailing List) - 149.36(if type: Loyalty Club Only) + 281.84(if type: Loyalty Club and Credit Card)

第 3 步：演示/可视化:

根据你的模型结果给出建议。（限 500 字以内）

根据在第 2 步得出的线性方程，

y = 303.46 + (66.98 * Avg Num Products Purchased) - 245.42(if type: Store Mailing List) - 149.36(if type: Loyalty Club Only) + 281.84(if type: Loyalty Club and Credit Card)

在“p1-mailinglist.xlsx”中，只采用以下的数值型自变量 Avg Num Products Purchased 和分类变量“Customer Segment”，并将“Customer Segment”做 3 个虚拟变量列即可。

通过上面的线性方程在“p1-mailinglist.xlsx“中求得预期盈利 y1。方程如下，

y1 = [303.46 + (66.98 * Avg Num Products Purchased) - 245.42(if type: Store Mailing List) - 149.36(if type: Loyalty Club Only) + 281.84(if type: Loyalty Club and Credit Card)]* Score_Yes – 6.5

$$\Sigma = 21987.96$$

即，若向 250 名新客户寄送产品目录，能给公司带来预期利润约为$21987.96。

至少回答以下问题：

1. 你的建议是什么？公司应该向这 250 个客户发送宣传册吗？
   我的建议是公司应该向这 250 个客户发送宣传册。
2. 你是如何得出你的建议的？（请解释你的推理流程，以便审核人员能够根据你的流程向你提供反馈）
   利用第 2 步得到的回归方程，在“p1-mailinglist.xlsx“中求得预期利润，
   预期利润乘以 Score_Yes，并减去 6.5 美元的产品目录册成本，就得到了单个预期利润，
   最后将单个客户的预期利润求和即可得到预期总利润$21987.96。
3. 新的宣传册带来的利润预计是多少？（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 21987.96 美元，远远大于决策条件里的 1 万美元。

报告人：王国瑞

2019.1.6