报告：预测宣传册需求

我们需要决定要不要向这 250 名新客户发放宣传册。

为了做出这个决策，我们需要获取的数据有：

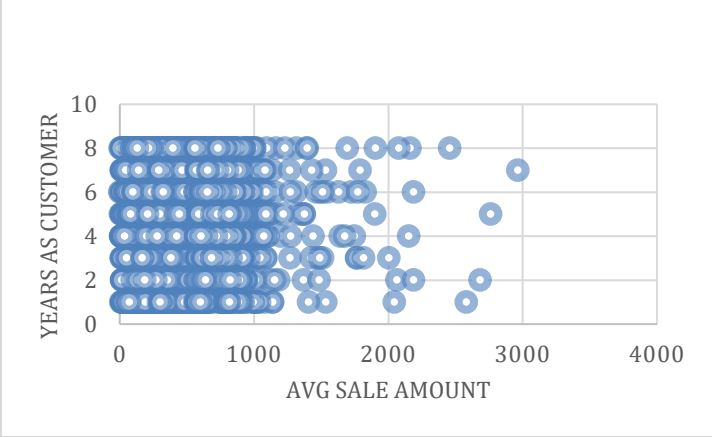
- 数据 1：已有客户的基本个体信息、购买信息、盈利额信息
- 数据 2：新客户的基本个体信息、购买信息
- 数据 3：新客户购买商品的概率
- 数据 4：产品的平均毛利率
- 数据 5：宣传册的成本

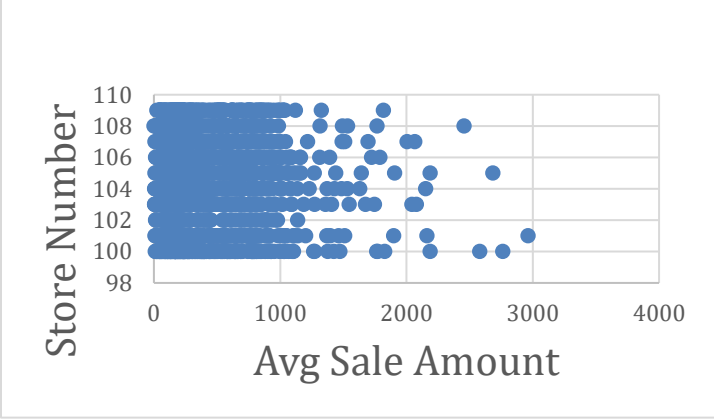
我们可以用这些数据预测新客户为公司带来的盈利额。如果盈利额预测超过一万美元，我们将对他们发放宣传册。

分析、建模和验证

首先，Avg Sale Amount 是线性回归 y 变量。

其次，根据常识，我认为 Name、Customer ID、Address、City、ZIP、Responded to Last Catalog 这些变量与盈利额相关度不大，不存在线性关系，予以删去。变量 State 所有数据相同，同样删去。Store Number、Years as Customer 这两个变量分别对 Avg Sale Amount 绘制散点图，散点图结果如下：





如图可见，都明显无线性关系。再用数据分析中的回归条件拟合，二者的 R 平方值都小于 0.001。因此，这两个变量也不予考虑。

在剩下的变量中，对于 Customer Segment 这一分类变量，以 Credit Card Only 为基本条件，对 Store Mailing List、Loyalty Club and Credit Card、Loyalty Club Only 设置虚拟变量。变量 Customer Segment 对 Avg Sale Amount 做线性回归:

SUMMARY OUTPUT

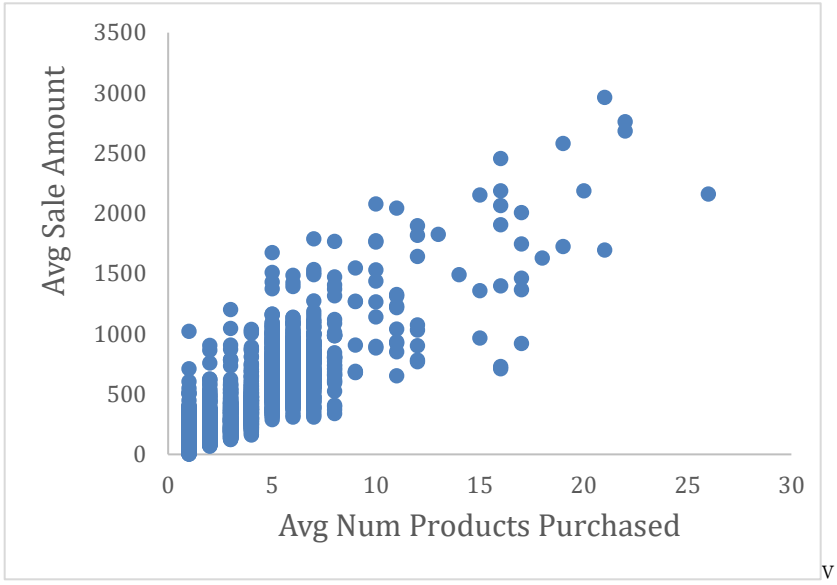
回归统计	
Multiple	0.838073
R Square	0.702367
Adjusted	0.70199
标准误差	185.6702
观测值	2375

方差分析

	df	SS	MS	F	gnificance 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			

	Coefficien	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
X Variabl	-525.317	10.04477	-52.2976	0	-545.015	-505.62	-545.015	-505.62
X Variabl	391.4805	15.73157	24.88503	1.2E-121	360.6315	422.3296	360.6315	422.3296
X Variabl	-286.346	11.37206	-25.1798	3.5E-124	-308.647	-264.046	-308.647	-264.046

对变量 Avg Num Products Purchased 和 Avg Sale Amount 绘制散点图，结果如下：



可以看出，两者近似有线性关系。线性回归拟合：

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%
Intercept	44.01516	5.704323	7.716107	1.75E-14	32.82919	55.20114	32.82919	55.20114
X Variabl	106.2802	1.319065	80.57237	0	103.6935	108.8668	103.6935	108.8668

由调整 R 平方和 p 值可以看出以上两种线性回归拟合良好，确定准备线性回归拟合的 y 值为变量 Avg Sale Amount，X1 为 Store Mailing List,X2 为 Loyalty Club and Credit Card，X3 为 Loyalty Club Only，X4 为 Avg Num Products Purchased。

最后，对三个变量进行线性回归拟合。得出结果如下：

SUMMARY OUTPUT

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

	Coefficien	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	282.7249	324.2021	282.7249	324.2021
X Variabl	-245.418	9.767776	-25.1252	1.1E-123	-264.572	-226.263	-264.572	-226.263
X Variabl	281.8388	11.90986	23.66433	2.6E-111	258.4839	305.1936	258.4839	305.1936
X Variabl	-149.356	8.972755	-16.6455	6.35E-59	-166.951	-131.76	-166.951	-131.76
X Variabl	66.9762	1.51504	44.20754	0	64.00526	69.94715	64.00526	69.94715

回归方程为 y=-245.42*X1+281.83*X2-149.36*X3 +66.98*X4+303.46

由调整 R 平方为 0.84，接近于 1，且截距和变量的 P 值分别为 1.1E-155、1.1E-123、2.6E-111、6.35E-59、0，都非常小，可知线性拟合良好。

综上所述，将线性回归方程结果用实际意义的变量表示如下：

Average Sale Amount=-245.42(If Type: Store Mailing List)+281.83(If Type: Loyalty Club and Credit Card)-149.36(If Type: Loyalty Club Only)+0(If Type: Credit Card Only)+66.98*Average Number Products Purchased+303.46

预测结果及建议

- 预测过程：

- 在新客户数据集中，对 Customer Segment 变量进行虚拟变量设置，然后通过拟合的线性回归模型，使用 Store Mailing List、Loyalty Club and Credit Card、Loyalty Club Only、Avg Num Products Purchased 几个变量代入进行预测，预测出每个客户的预测盈利额。
- 将每个客户的预测盈利额乘以购买概率，再乘以产品的平均毛利率，最后减去宣传册的成本，得出每个客户带给公司的的实际盈利。
- 将所用客户盈利额求和，得出公司的实际总营业额数据。
    - 最终预测盈利额

公司实际收获的利润是 21987.91 美元

    - 建议

根据预测结果，如果向新客户群体发放宣传册，实际盈利将超过经理要求的一万美元。因此，我建议公司向这 250 个客户发送宣传册。

(参考资料：N/A)