# 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？

   需要向 250 位目标客户寄送目录册

2. 做出这些决策需要获取哪些数据？

| 数据项 | 数据名称 | 数据来源 | 解释 |
|---|---|---|---|
| **1.** | Avg Sale Amount | p1-customers | 在建模过程中作为目标变量，拟合函数 |
| **2** | Avg Num Products purchased | p1-customers | 在建模过程中作为预测变量，拟合函数 |
| **3** | # Years as Customer | p1-customers | 在建模过程中作为预测变量，拟合函数 |
| **4** | Customer Segment | p1-customers | 在建模过程中建立虚拟变量，拟合函数 |
| **5** | Responded to Last Catalog | p1-customers | 在建模过程中作为预测变量，拟合函数 |
| **6** | Avg Num Products Purchased | p1-mailinglist | 用于带入函数，从而得出顾客的消费金额 |
| **7** | # Years as Customer | p1-mailinglist | 用于带入函数，从而得出顾客的消费金额 |
| **8** | Customer Segment | p1-mailinglist | 用于带入函数，从而得出顾客的消费金额 |

| | | | |
|---|---|---|---|
| **9** | Score_Yes | p1-mailinglist | 用于带入函数，从而得出顾客的消费金额 |
| **10** | **产品目录册的成本** | 已知数据 | 用于计算利润 |
| **11** | **毛利率** | 已知数据 | 用于计算利润 |

# 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）。*

*重要事项：使用* ***p1-customers.xlsx 训练你的线性模型。***
*至少回答以下问题：*

1. 你是如何在你的模型中选择预测变量（请参阅补充文本）的？原因是什么？你必须解释你选择的连续预测变量与目标变量有线性关系。请参阅这节课来探索你的数据，并使用散点图寻找线性关系。你必须在答案中包含散点图。
   目标变量：平均销售金额（Avg Sale Amount）
   预测变量：
   预测变量 1：平均购买件数（Avg Num Products purchased）
   预测变量 2：成为顾客的时间(Years as customers)
   预测变量 3：顾客群---是否为 Store Mailing List;
   预测变量 4：顾客群---是否为 Loyalty Club and Credit Card;
   预测变量 5：顾客群---是否为 Loyalty Club only;
   预测变量 6：上次是否回复宣传册（Responded to Last Catalog）

1.1 平均购买件数（Avg Num Products purchased）与平均销售金额（Avg Sale Amount）的关系，P=0，小于 0.05,说明相关性强，可以作为预测变量

SUMMARY OUTPUT

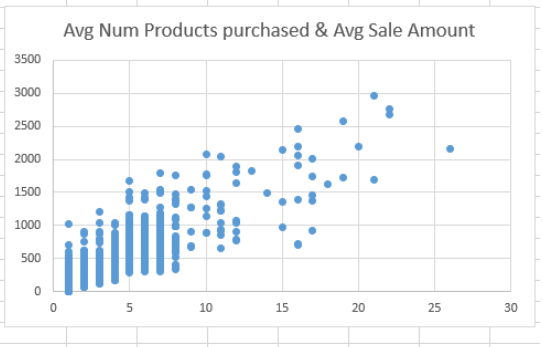
回归统计

| | |
|---|---|
| Multiple | 0.85575 |
| R Square | 0.73232 |
| Adjustec | 0.7322 |
| 标准误差 | 176.007 |
| 观测值 | 2375 |

方差分析

| | df | SS | MS | F | nificance F |
|---|---|---|---|---|---|
| 回归分析 | 1 | 2E+08 | 2E+08 | 6491.91 | 0 |
| 残差 | 2373 | 7.4E+07 | 30978.5 | | |
| 总计 | 2374 | 2.7E+08 | | | |

| | Coefficien | 标准误差 | t Stat | P-value | ower 95% | pper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercep | 44.0152 | 5.70432 | 7.71611 | 1.8E-14 | 32.8292 | 55.2011 | 32.8292 | 55.2011 |
| X Variab | 106.28 | 1.31906 | 80.5724 | 0 | 103.694 | 108.867 | 103.694 | 108.867 |



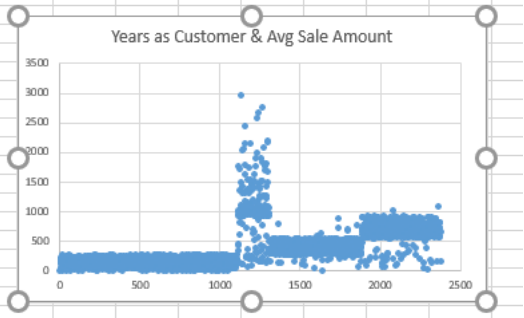

1.2 成为顾客的时间(Years as customers)与平均销售金额（Avg Sale Amount）的关系，P 值大于 0.05,说明相关性弱

SUMMARY OUTPUT

| 回归统计 | |
|---|---|
| Multipl | 0.02978 |
| R Squar | 0.00089 |
| Adjuste | 0.00047 |
| 标准误差 | 340.037 |
| 观测值 | 2375 |

| 方差分析 | df | SS | MS | F | nificance F |
|---|---|---|---|---|---|
| 回归分析 | 1 | 243578 | 243578 | 2.10662 | 0.14679 |
| 残差 | 2373 | 2.7E+08 | 115625 | | |
| 总计 | 2374 | 2.7E+08 | | | |

| | Coefficien | 标准误差 | t Stat | P-value | ower 95% | pper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Interce | 380.039 | 15.2829 | 24.8669 | 2E-121 | 350.07 | 410.008 | 350.07 | 410.008 |
| X Varia | 4.385 | 3.02118 | 1.45142 | 0.14679 | -1.5394 | 10.3094 | -1.5394 | 10.3094 |




1.3 将顾客群（Customer Segement）类别变量转化为虚拟变量，Store Mailing List，Loyalty Club and Credit Card，Loyalty Club Only

顾客群（Customer Segement）与平均销售金额（Avg Sale Amount）的关系，

Store Mailing List 中 P=0 小于 0.05,说明相关性强

Loyalty Club and Credit Card，Loyalty Club Only 中，P=0 大于 0.05,说明相关性弱

总统上，顾客群（Customer Segement）与平均销售金额（Avg Sale Amount）的关系存在一定相关性，说明顾客群（Customer Segement）可以作为预测变量

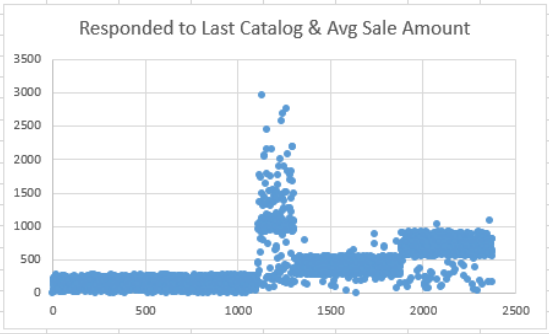
1.4 上次是否回复宣传册（Responded to Last Catalog）与平均销售金额（Avg Sale Amount）的关系，P=0 大于 0.05,说明相关性弱

SUMMARY OUTPUT

| 回归统计 | |
|---|---|
| Multiple | 0.00575 |
| R Square | 3.3E-05 |
| Adjusted | -0.0004 |
| 标准误差 | 340.182 |
| 观测值 | 2375 |

| 方差分析 | df | SS | MS | F | nificance F |
|---|---|---|---|---|---|
| 回归分析 | 1 | 9068.52 | 9068.52 | 0.07836 | 0.77955 |
| 残差 | 2373 | 2.7E+08 | 115724 | | |
| 总计 | 2374 | 2.7E+08 | | | |

| | Coefficien | 标准误差 | t Stat | P-value | Lower 95% | Upper 95% | 下限 95.0% | 上限 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercep | 400.884 | 8.02709 | 49.9414 | 0 | 385.143 | 416.624 | 385.143 | 416.624 |
| X Variab | -4.551 | 16.2574 | -0.2799 | 0.77955 | -36.431 | 27.3291 | -36.431 | 27.3291 |




综上，只有平均购买件数（Avg Num Products purchased）和顾客群（Customer Segement）与平均销售金额（Avg Sale Amount）具有较强的相关性，用于拟合回归函数。

2. 解释为何你认为你的线性模型是很好的模型。必须使用你的回归模型产生的统计学结果证明你的推理过程。对于你所选择的每个变量，请使用你的模型产生的 p 值和 R 平方值证明每个变量为何与你的模型很好地拟合。

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% |
|---|---|---|---|---|---|---|---|---|
| Intercep | 303.4635 | 10.57571 | 28.69437 | 1.1E-155 | 282.7249 | 324.2021 | 282.7249 | 324.2021 |
| X Variab | 66.9762 | 1.51504 | 44.20754 | 0 | 64.00526 | 69.94715 | 64.00526 | 69.94715 |
| X Variab | -245.418 | 9.767776 | -25.1252 | 1.1E-123 | -264.572 | -226.263 | -264.572 | -226.263 |
| X Variab | 281.8388 | 11.90986 | 23.66433 | 2.6E-111 | 258.4839 | 305.1936 | 258.4839 | 305.1936 |
| X Variab | -149.356 | 8.972755 | -16.6455 | 6.35E-59 | -166.951 | -131.76 | -166.951 | -131.76 |

由回归分析表可知，$R^2$的约值为 0.84，值接近 1，说明回归直线对观测值的拟合程度越好说明很大程度可以解释预测变量导致目标变量的改变。
总体来说整个回归方程呈现出正相关的趋势，说明此回归模型有很好地拟合。

3. 最佳线性回归方程

Y = 303.46+ 66.98* Avg Num Products purchased– 245.42(If Type: Store Mailing List) + 281.84(If Type: Loyalty Club and Credit Card) – 149.37(If Type: Loyalty Club only)+ 0(If Type: Credit Card only)

# 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）*

*至少回答以下问题：*

1. 我的建议：公司应该向这 250 个客户发送宣传册
2. 原因：利润值 > 10000 美元，即满足经理的要求可以向客户发送目录册。
   预计新的宣传册带来的利润预计：销售量* (store_yes) * 50% – $6.5
   经计算：
   总利润 = $21987.75
3. 所以公司应该向这 250 个客户发送宣传册