预测宣传册需求

## 第 1 步：理解业务和数据

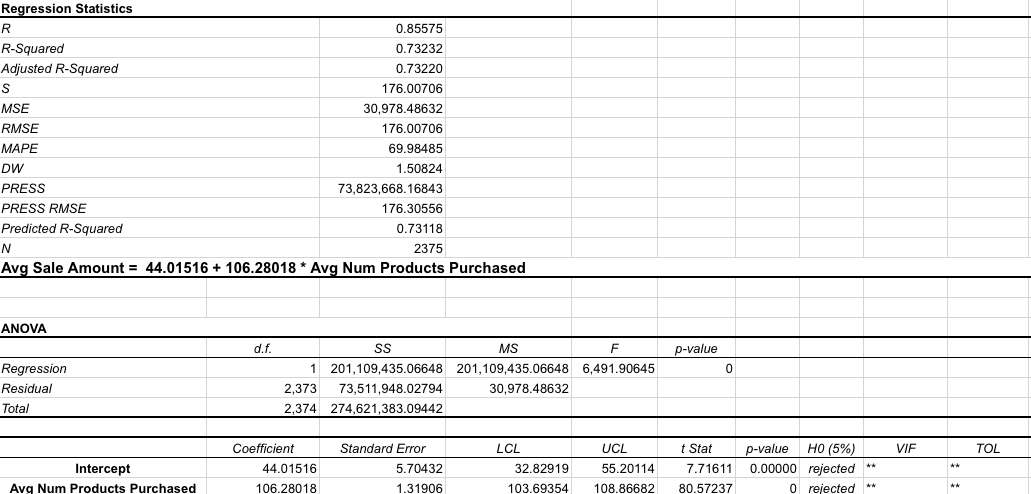
我们需要作出的决策是：是否需要向250名新客户寄送产品目录册？

作出决策所需的数据：

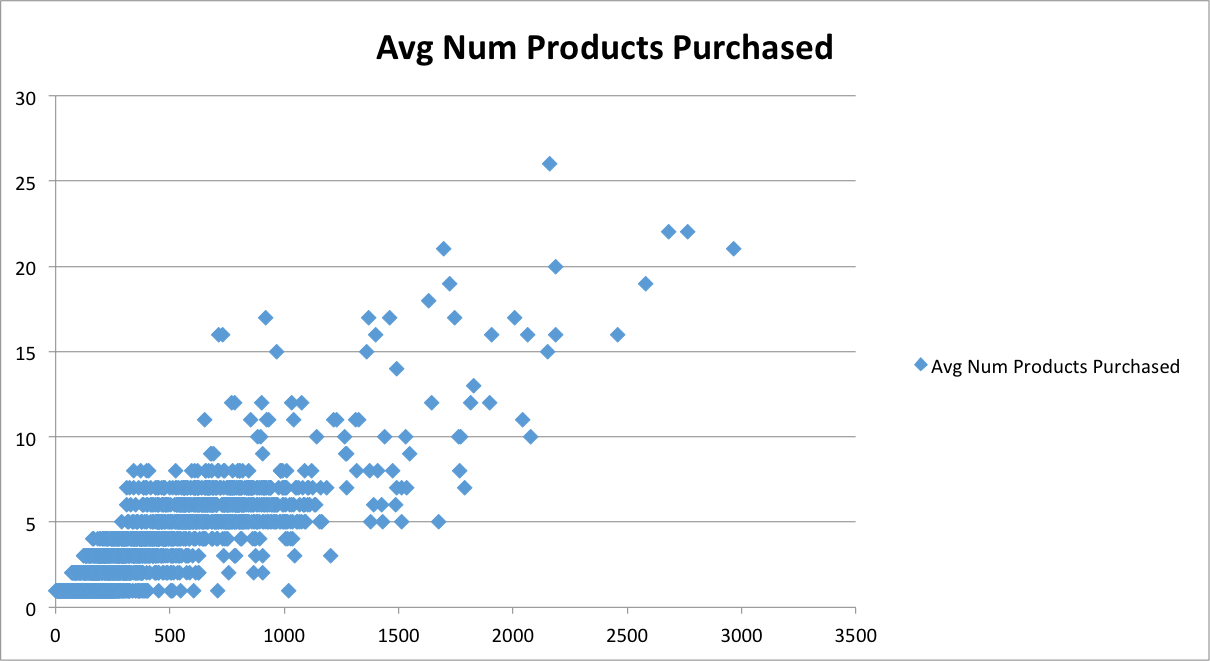
|  |  |  |  |
| --- | --- | --- | --- |
| 数据项 | 数据名称 | 数据来源 | （进一步）解释 |
| 1 | avg number products purchased | P1-customers.xlsx | 在建模过程中作为预测变量 |
| 2 | customer segment | P1-customers.xlsx | 在建模过程中作为虚拟变量 |
| 3 | avg sale amount | P1-customers.xlsx | 在建模过程中作为目标变量 |
| 4 | customer segment | P2-mailing list.xlsx | 代入线性回归方程模型作为预测变量 |
| 5 | avg number products purchased | P2-mailing list.xlsx | 代入线性回归方程模型作为预测变量 |
| 6 | Score\_Yes | P2-mailing list.xlsx | 需要在计算收益时作为顾客的购买概率考虑进去 |
| 7 | 印刷和寄送目录册的成本 | 项目辅助材料 | 需要作为成本在计算收益时考虑去除 |
| 8 | 产品的毛利率 | 项目辅助材料 | 需要在计算收益时考虑进去 |

## 第 2 步：分析、建模和验证

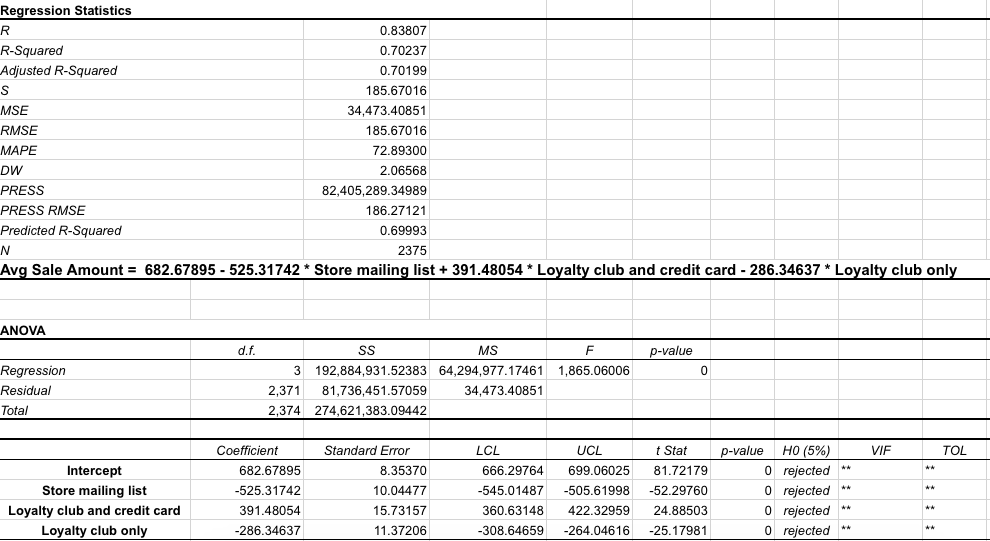
首先选择变量avg number products purchased，预测变量和目标变量avg sale amount之间的线性关系如下，可以看出R平方>0.7，属于强相关，同时P值为0，说明其与目标标量之间的关系具有统计学意义



经过制作散点图，可以看出和目标变量avg sale amount之间有很明显的线性关系，如图所示：



对于变量customer segment，以credit card only作为基础条件，其它三个预测变量和目标变量之间的线性关系如下，可以看出R平方>0.7，属于强相关，同时P值为0，说明其与目标标量之间的关系具有统计学意义



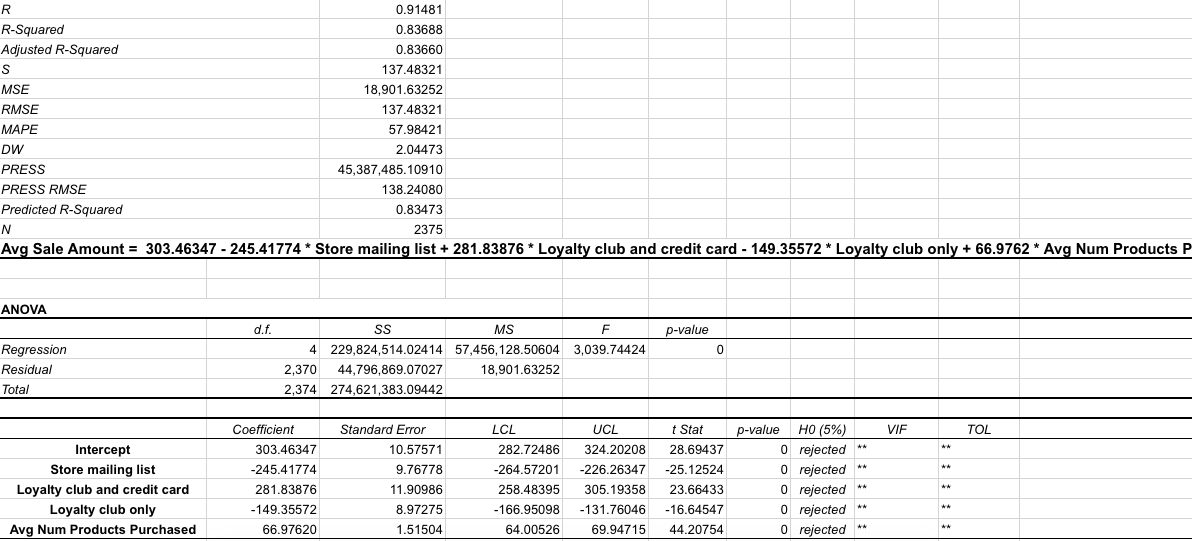
所以我们最终的线性模型主要与avg number products purchased和customer segment两个预测变量相关，具体的线性方程为：

根据原始数据建立起的线性方程模型：

Y=303.46-245.42(If Type: Store Mailing List)+281.84(If Type: Loyalty Club and Credit Card)-149.36(If Type: Loyalty Club Only)+66.98\*Avg Num Products Purchased

最终的线性方程模型：

Y=(303.46-245.42(If Type: Store Mailing List)+281.84(If Type: Loyalty Club and Credit Card)-149.36(If Type: Loyalty Club Only)+66.98\*Avg Num Products Purchased)\*Score\_Yes\*50%-6.5



线性方程中设置虚拟变量customer segment中的credit card only为基准值，那么store mailing list的系数指在avg num products purchased相同的情况下，我们预计store mailing list的avg sale amount比credit card only的收入要少245.42美元；同样的，loyalty club and credit card的系数指在avg num products purchased相同的情况下，预计loyalty club and credit card的avg sale amount比credit card only的收入要多281.84美元；loyalty club only的系数指在avg num products purchased相同的情况下，预计loyalty club only的avg sale amount比credit card only的收入要少149.36美元。

avg num products purchased的系数指在其它变量一定的情况下，多消费一件物品，平均销售额上涨66.98美元。

线性方程的R平方值>0.7，是强相关模型，四个预测变量的p值都是0，说明其和目标变量之间的关系具体显著的统计学关系。

## 第 3 步：演示/可视化:

根据数据分析所建立起的线性回归方程模型，可代入250个新客户的数据资料进行计算，最终所得的预测收益为21987.96美元，高于所要达到的目标10000美元，故应该向这250个客户发送宣传册。

1. 当新客户的customer segment为store mailing list时，根据线性回归方程模型进行计算，所得收益为529.68美元；
2. 当新客户的customer segment为loyalty club and credit card时，根据线性回归方程模型进行计算，所得收益为4602.72美元；
3. 当新客户的customer segment为loyalty club only时，根据线性回归方程模型进行计算，所得收益为7849.89美元；
4. 当新客户的customer segment为credit card only时，根据线性回归方程模型进行计算，所得收益为9005.67美元；

四项合计总值为21987.96美元，而当新客户带来的预期利润超过10000美元时，公司就会选择向250名新客户发送宣传册，故建议公司向这250名客户发送宣传册。