预测宣传册需求

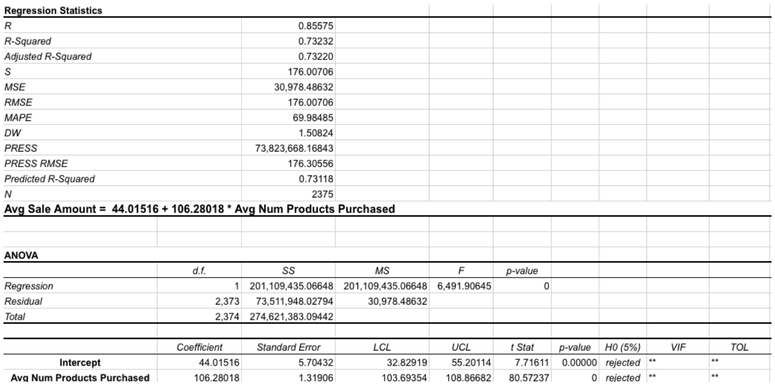
第 1 步：理解业务和数据

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



Regression Statistics

R	0.85575
R-Squared	0.73232
Adjusted R-Squared	0.73220
S	176.00706
MSE	30,978.48632
RMSE	176.00706
MAPE	69.98485
DW	1.50824
PRESS	73,823,668.16843
PRESS RMSE	176.30556
Predicted R-Squared	0.73118
N	2375

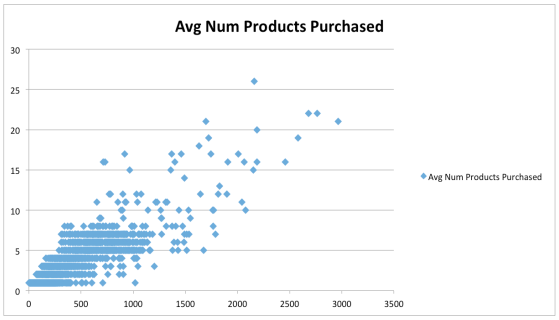
Avg Sale Amount = 44.01516 + 106.28018 * Avg Num Products Purchased

ANOVA

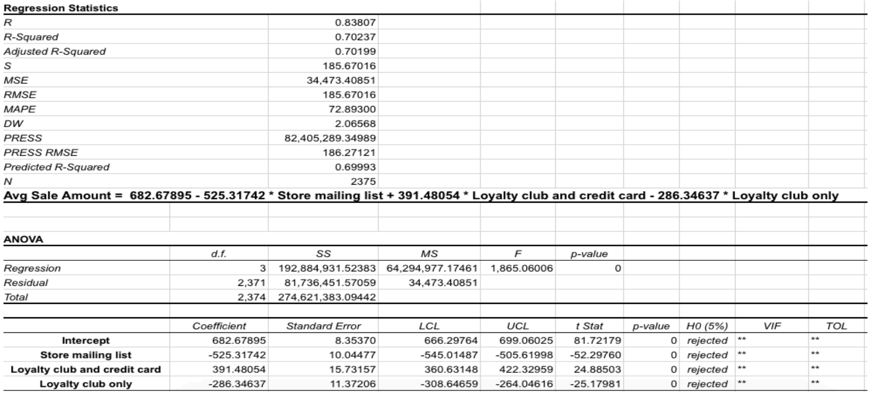
	d.f.	*SS*	*MS*	*F*	*p-value*
Regression	1	201,109,435.06648	201,109,435.06648	6,491.90645	0
Residual	2,373	73,511,948.02794	30,978.48632		
Total	2,374	274,621,383.09442			

	Coefficient	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-value*	*H0 (5%)*	*VIF*	*TOL*
Intercept	44.01516	5.70432	32.82919	55.20114	7.71611	0.00000	*rejected*	**	**
Avg Num Products Purchased	106.28018	1.31906	103.69354	108.86682	80.57237	0	*rejected*	**	**

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



Regression Statistics

R	0.83807
R-Squared	0.70237
Adjusted R-Squared	0.70199
S	185.67016
MSE	34,473.40851
RMSE	185.67016
MAPE	72.89300
DW	2.06568
PRESS	82,405,289.34989
PRESS RMSE	186.27121
Predicted R-Squared	0.69993
N	2375

Avg Sale Amount = 682.67895 - 525.31742 * Store mailing list + 391.48054 * Loyalty club and credit card - 286.34637 * Loyalty club only

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-value*
Regression	3	192,884,931.52383	64,294,977.17461	1,865.06006	0
Residual	2,371	81,736,451.57059	34,473.40851		
Total	2,374	274,621,383.09442			

	Coefficient	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-value*	*H0 (5%)*	*VIF*	*TOL*
Intercept	682.67895	8.35370	666.29764	699.06025	81.72179	0	*rejected*	**	**
Store mailing list	-525.31742	10.04477	-545.01487	-505.61998	-52.29760	0	*rejected*	**	**
Loyalty club and credit card	391.48054	15.73157	360.63148	422.32959	24.88503	0	*rejected*	**	**
Loyalty club only	-286.34637	11.37206	-308.64659	-264.04616	-25.17981	0	*rejected*	**	**

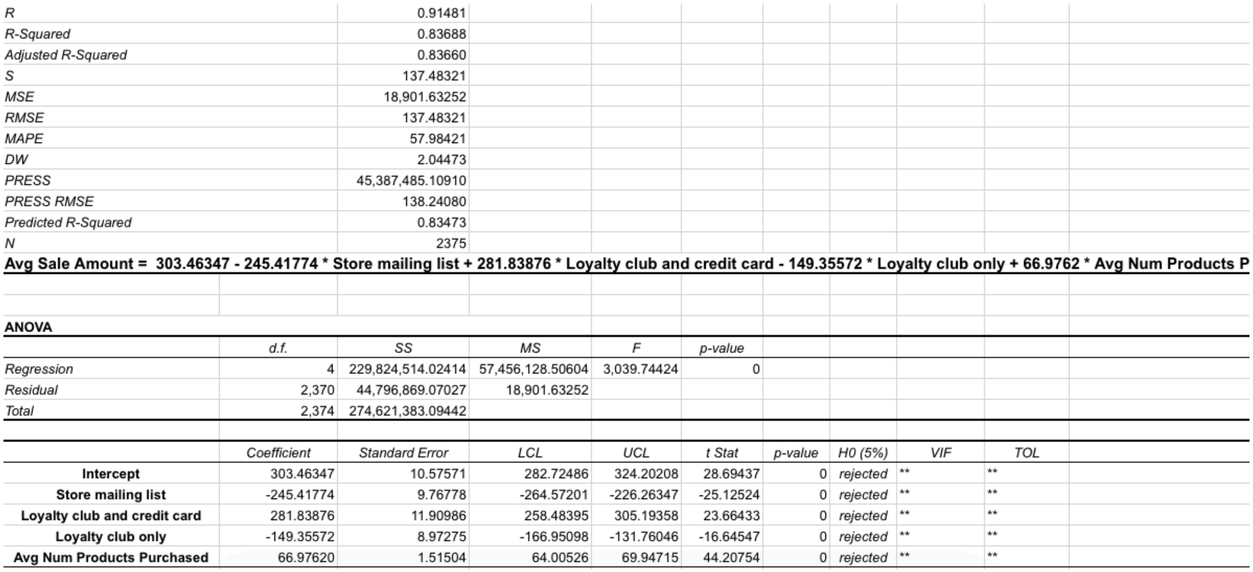
所以我们最终的线性模型主要与 avg number products purchased 和 customer segment 两个预测变量相关，具体的线性方程为：

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



R	0.91481
R-Squared	0.83688
Adjusted R-Squared	0.83660
S	137.48321
MSE	18,901.63252
RMSE	137.48321
MAPE	57.98421
DW	2.04473
PRESS	45,387,485.10910
PRESS RMSE	138.24080
Predicted R-Squared	0.83473
N	2375

Avg Sale Amount = 303.46347 - 245.41774 * Store mailing list + 281.83876 * Loyalty club and credit card - 149.35572 * Loyalty club only + 66.9762 * Avg Num Products P

ANOVA

	d.f.	SS	MS	F	p-value
Regression	4	229,824,514.02414	57,456,128.50604	3,039.74424	0
Residual	2,370	44,796,869.07027	18,901.63252		
Total	2,374	274,621,383.09442			

	Coefficient	Standard Error	LCL	UCL	t Stat	p-value	H0 (5%)	VIF	TOL
Intercept	303.46347	10.57571	282.72486	324.20208	28.69437	0	rejected	**	**
Store mailing list	-245.41774	9.76778	-264.57201	-226.26347	-25.12524	0	rejected	**	**
Loyalty club and credit card	281.83876	11.90986	258.48395	305.19358	23.66433	0	rejected	**	**
Loyalty club only	-149.35572	8.97275	-166.95098	-131.76046	-16.64547	0	rejected	**	**
Avg Num Products Purchased	66.97620	1.51504	64.00526	69.94715	44.20754	0	rejected	**	**

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。