预测宣传册需求

请完成每个部分。准备好后，将你的文件另存为 PDF 文档并从课堂上提交。

v3 更新：

1. 非常感谢老师的迅速反馈。
2. 解释的部分明白了，非常清楚。
3. 修改部分已完成。

v2 更新：

1. 感谢评审老师在新年这天的光速回复！
2. 两个问题都做了修改，如黄色底色的更新。
3. 对于反馈修改公式中的：因为此处的虚拟变量实际并不是与系数相乘的关系，而是与其余几个虚拟变量互斥的条件关系。可不可以这样理解**问题 1（答复：正确）**：
   a) 虽然是采用乘的方式计算，但实际上这几个虚拟变量间是排除关系。
   b) 这种排除关系是当标准变量（舍弃的那个，3 个分类变量没有调整，为 0）
   c) 当变量是 3 个虚拟变量中的一个时，对结果做修正，增加的是系数那么多。
4. 另外，更新时又有了个新的问题想请教**问题 2**:
   a) 最后计算时，对于决定系数 Adjust R-squre 和 P 同时做比较，遇到的情况是两个都是同向的变化。
   b) 从概念理解看，P 是单个 Feature 的衡量指标，而 Adjust R-squre 是算法的衡量指标。
   c) 那么想请教这两个值有没有换算关系？并且会不会出现两个指标冲突的情况？
   d) 答复：p 值和 r 平方值间没有换算关系，两个值之间也没有必要联系。当 p 值小且 r 平方值也小的情况可能是其中某个参与分析的变量相关性不大。谨记 p 值是变量与预测变量间线性关系是否显著的判断依据，r 平方值检验的是整个线性模型的效力。

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）

关键决策：

请回答以下问题

1. 需要作出什么样的决策？
   a. 针对 mailing list 中的 250 个人，如果利润超过 1 万美元，我们就寄出产品手册。
   b. （因为每个用户都会有估算，根据计算出的排序也可以比较灵活）
   c. （好厚好厚的一本，成本可贵了，购买意愿低的就算了）

d. （也就是说可以给邮件目录中利润大于成本的人寄，利润小于成本的就放弃）

2. 作出这些决策需要获取哪些数据？（根据 customers 文件中的记录，又些做了清理）

a. 比如 customer 这列是用户 id，丢弃。
b. states 这列都一样（数据来自同一个州），丢弃。
c. city 和 store number 没有对应关系，检查了 mailing list 也有这个输入，所以两个都保留（同一个 store number 在好几个城市都有出现）。后续操作的时候发现 city 需要做的虚拟变量太多了（偷懒）放弃 City，保留 Store。
d. name 列比较独特，和 customer id 类似，丢弃。
e. 补充，数据字典：

数据项	数据名称	数据来源	说明
1	Store Number	P1-customers.xlsx P1-mailinglist.xlsx	每个城市的商店编号，分类数据。需用虚拟变量处理
2	Average Sale	P1-customers.xlsx	要预测的变量，数值数据。
3	Customer Segment	P1-customers.xlsx P1-mailinglist.xlsx	客户分类变量，分类数据。需用虚拟变量处理
4	Years as Customer	P1-customers.xlsx P1-mailinglist.xlsx	客户时长变量，分类数据（离散）。
5	City	P1-customers.xlsx P1-mailinglist.xlsx	城市变量，分类数据。需用虚拟变量处理
6	Cost per Catalog	P1-mailinglist.xlsx	6.5 each，根据描述加入到文件中进行计算
7	Benefit Ratio	P1-mailinglist.xlsx	50%，根据描述加入到文件中进行计算

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。

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。
3. 根据提供的数据，最佳线性回归方程是什么？每个系数小数点后最多保留两位（例如 1.28）

   重要事项：回归方程应该为以下形式
   *Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

   例如：Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

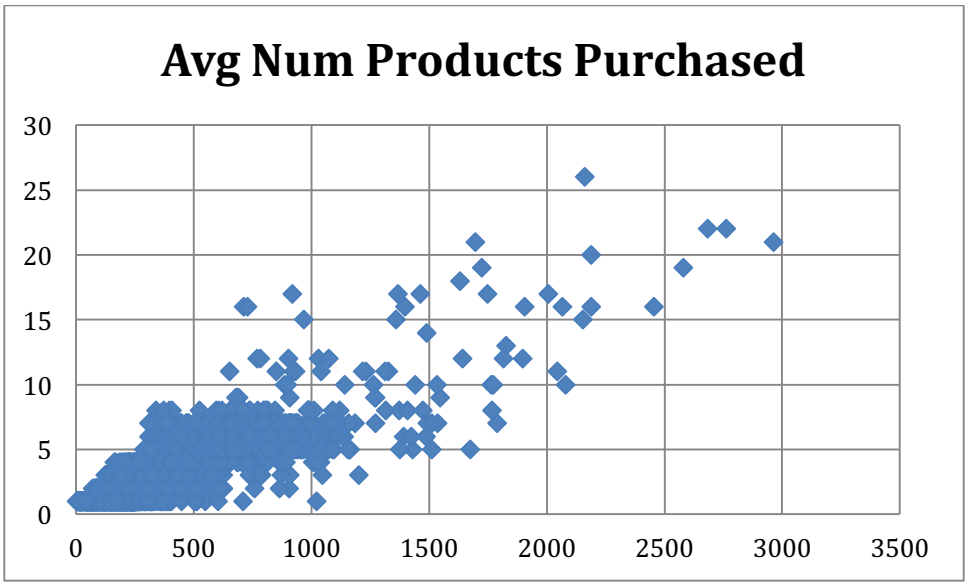
   注意，对于类型 Cash ，我们**必须**包含系数 0。

注意：如果你使用的是 Alteryx 之外的其他软件，并且决定使用 Customer Segment 作为其中一个预测变量，请将基本条件设为 Only Credit Card。

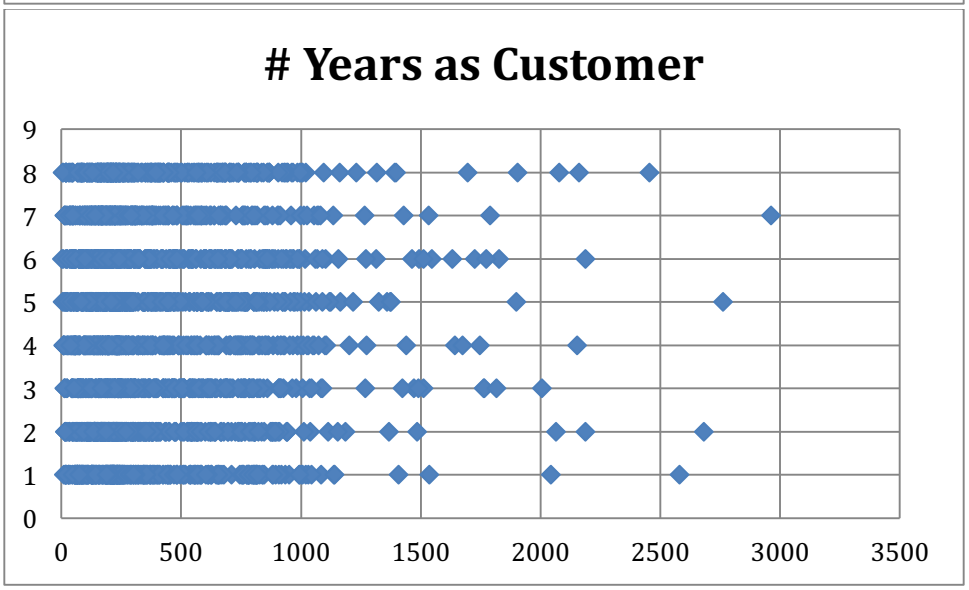
问题 1 回答：

1. 变量的选择是
   a) responded 是在寄出之后才能得到的数据，用作后续分析，丢弃。
   b) 清理后的原始数据为：Customer Segment、Avg Sale、Store Num。
   c) 因变量是 Avg Sale，其他的是自变量。并且全部需要进行虚拟变量的转换。
   d) city 和 store number 没有对应关系，检查了 mailing list 也有这个输入，所以两个都保留（同一个 store number 在好几个城市都有出现）。后续操作的时候发现 city 需要做的虚拟变量太多了（偷懒）放弃 City，保留 Store。
   e) 从数据观察（描述统计学）来观察，只能画出 2 个散点图（Avg Sale 与 Avg Num，Avg Sale 与 Years as Customer），剩下的分类信息需要在回归部分计算。从两个散点图观察，前一个有线性相关性，后一个没有线性相关性，两个图如下（抱歉 Mac 的散点图无法画出预测线）：

f)



g)



问题 2 解答：

1. 线性模型算法在这个项目因为我们要进行预测的是每个客户的消费额。
2. 现在对于选中的几个变量分别做线性回归，检查 Adjust R-Squre，结果如下：
   a) Avg Num 值为 0.732，p 接近 0，入选。
   b) Years as Customer 值为 0.004 ， p=0.147（达不到统计显著性）放弃。
   c) Customer Segment 值为 0.702，p 接近 0，入选。
   d) Store Number 值为 -0.006，放弃，p 大于 0.05，放弃。
   e) 请教 1：为什么 Store Number 的输出会是负的，输出如下：（答复：我们不用特别在意系数是正或负，拟合过程中的系数正负由变量本身与预测变量决定。）

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*回归统计*					
4	Multiple R	0.05616378				
5	R Square	0.00315437				
6	Adjusted R Sc	-0.0006391				
7	标准误差	340.224479				
8	观测值	2375				
9						
10	方差分析					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	回归分析	9	866257.512	96250.8347	0.83152132	0.58697183
13	残差	2365	273755126	115752.696		
14	总计	2374	274621383			
15						
16		*Coefficients*	*标准误差*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	427.162663	26.1711137	16.3219138	7.836E-57	375.841958
18	X Variable 1	-14.657172	32.2489816	-0.4545003	0.64951048	-77.896379
19	X Variable 2	-29.69875	33.2313259	-0.8936974	0.37157475	-94.864302
20	X Variable 3	-46.734663	45.2407296	-1.0330219	0.30169929	-135.45027
21	X Variable 4	-20.669463	34.6321183	-0.5968293	0.55067849	-88.581923
22	X Variable 5	-40.889811	33.3712583	-1.2253002	0.22058413	-106.32977
23	X Variable 6	-7.9218103	32.6258215	-0.242808	0.8081752	-71.899988
24	X Variable 7	-45.140684	33.0748745	-1.3648029	0.17244484	-109.99944
25	X Variable 8	-13.159742	34.5992379	-0.3803478	0.70372145	-81.007726
26	X Variable 9	-67.830329	35.1586477	-1.9292645	0.05381755	-136.7753

f)

g) **请教 2：（例如上例）如果输出结果的多个 Variable 的 P-value 有显著的，也有不显著的（大于 0.05 的和小于 0.05 的都有），那么该怎么理解和解释呢？（答复：非常棒的观察！因为 store number 为分类变量，其中的各项值可能存在大于显著标准的可能，如 customer segment 中的分类变量也存在这样的情况。一般来说，如果大多数变量均有显著性便可添加入多元变量拟合中，存在一个或两个变量 p 值稍微大一些也可以，不过在报告中需指明。但是 store number 中大多数值的 p 值不满足 0.05 这个门框，所以不用考虑 store number。）**

问题 3 解答：

1. 根据问题 2 的答案，我们得出的结果是：

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*回归统计*								
4	Multiple R	0.9148102							
5	R Square	0.83687771							
6	Adjusted R Sc	0.8366024							
7	标准误差	137.483208							
8	观测值	2375							
9									
10	方差分析								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	回归分析	4	229824514	57456128.5	3039.74424	0			
13	残差	2370	44796869.1	18901.6325					
14	总计	2374	274621383						
15									
16		*Coefficients*	*标准误差*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*下限 95.0%*	*上限 95.0%*
17	Intercept	303.463471	10.5757148	28.6943697	1.123E-155	282.72486	324.202083	282.72486	324.202083
18	X Variable 1	66.9762049	1.51504036	44.2075385	0	64.0052631	69.9471467	64.0052631	69.9471467
19	X Variable 2	-245.41774	9.76777562	-25.125244	1.05E-123	-264.57201	-226.26347	-264.57201	-226.26347
20	X Variable 3	281.838765	11.9098574	23.6643274	2.58E-111	258.483946	305.193584	258.483946	305.193584
21	X Variable 4	-149.35572	8.97275479	-16.64547	6.3458E-59	-166.95098	-131.76046	-166.95098	-131.76046

2.

3. 回归方程为：Y = 303.46 + 66.98Average_Number_of_Buying + -245.42(if Type:Store Mailing List) + 281.84(if Type:Loyalty Club and Credit Card) + 0 (If Type: Credit Card Only)

4. *在最后，我还和对使用 4 个 feature 的回归方程结果做了对比，结果只在小数点后地 4 位开始有微小区别：*

回归统计	
Multiple R	0.91523457
R Square	0.83765432
Adjusted R Sc	0.83669126
标准误差	137.44582
观测值	2375

5.

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. 你的建议是什么？公司应该向这 250 个客户发送宣传册吗？

答：我们要向这 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）

答：推理流程如下。

a) 根据之前的数据计算，我们发现有两个自变量和客户产生的利润：客户的付款方式与客户一次购买的商品数量。

b) 分析得出的多元线性回归方程是：Y = 303.46 + 66.98Average_Number_of_Buying + -245.42(if Type:Store Mailing List) + 281.84(if Type:Loyalty Club and Credit Card) + 0 (If Type: Credit Card Only)

c)

d) 将产品手册 Cost=6.5 和 Ratio=0.5 带入计算，得出结果比预期高很多。

e) 详细计算截图如下：

f)

Y	Predict_Sale	benefit_finial	Cost	Ratio	SUM		
355.04	108.2999129	47.6499565	6.5	0.5	21987.95703	equation:	
987.18	466.6642084	226.832104					value
622.96	360.6202373	173.810119				w1	66.98
288.06	87.89799784	37.4489989				w2	-245.42
422.02	163.6196249	75.3098125				w3	281.84
772.32	206.4243712	96.7121856				w4	-149.36
853.22	189.1925711	88.0962855				intercept	303.46

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）

答：利润是 21988 美元，超过 10000 美元 1 倍多。

提交之前

请根据此处的审核标准中列出的项目要求检查你的答案。审核人员将根据该审核标准对项目打分。