Passo 1: Compreensão do Negócio e dos Dados

Decisões Chaves:

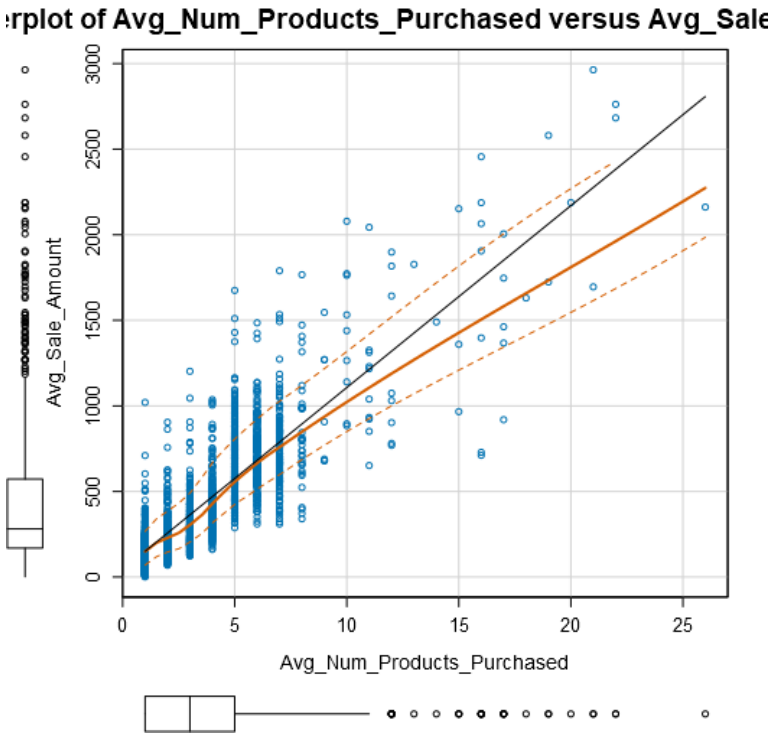
Responda estas perguntas

1. Que decisões precisam ser feitas??
   Se os catálogos devem ser enviados à lista de 250 contatos.

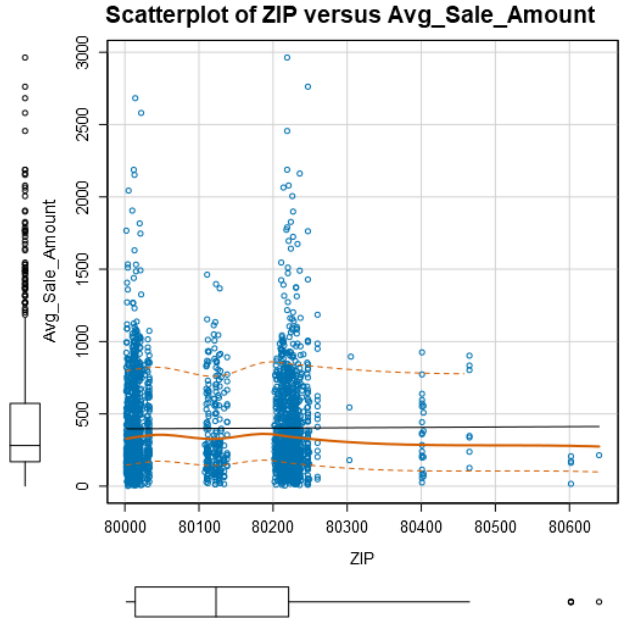
2. Que dados são necessários para subsidiar essas decisões??
   Qual o lucro esperado, probabilidade de compra de cada consumidor e os custos associados à campanha.

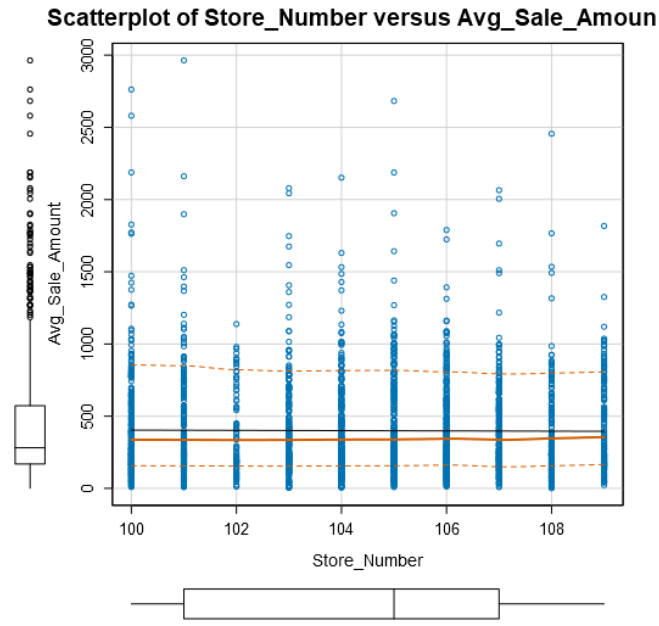
Passo 2: Análise, modelagem e validação

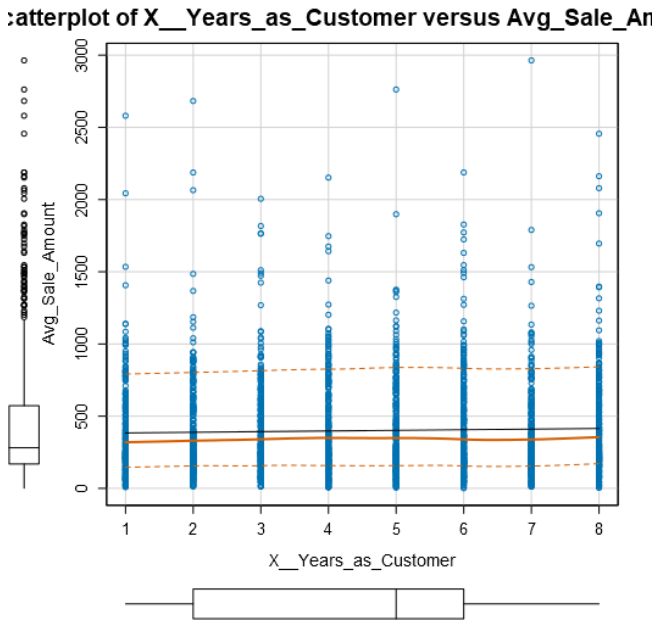
Inicialmente foram plotados gráficos de dispersão para todas as variáveis não categóricas com o fim de verificar a existência de relações lineares entre as variáveis independentes e a variável dependente, com apenas a variável *Avg Num Products Purchased* demonstrando tal linearidade. O resultado das outras variáveis pode ser visto abaixo.




Scatterplot of Customer_ID versus Avg_Sale_Amount
Avg_Sale_Amount
Customer_ID
0
500
1000
1500
2000
2500
3000


Scatterplot of ZIP versus Avg_Sale_Amount
Avg_Sale_Amount
ZIP
80000
80100
80200
80300
80400
80500
80600


Scatterplot of Store_Number versus Avg_Sale_Amoun
Avg_Sale_Amount
Store_Number
100
102
104
106
108


catterplot of X__Years_as_Customer versus Avg_Sale_An
Avg_Sale_Amount
X__Years_as_Customer
1
2
3
4
5
6
7
8

Visando observar a significância das variáveis categóricas e confirmação das variáveis contínuas, foi rodado um modelo preliminar. Tal modelo foi capaz de confirmar a relevância da variável *Avg Num Products Purchased* e indicou a relevância da variável categórica *Customer Segment*, que apresentaram valor-p bem abaixo das outras.

Report

Report for Linear Model Linear_Regression_37

Basic Summary

Call:

lm(formula = Avg.Sale.Amount ~ Customer.Segment + City + ZIP + Store.Number + Avg.Num.Products.Purchased + X..Years.as.Customer, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-678.60	-65.28	-1.56	70.86	959.00

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.189e+04	9505.8914	2.302956	0.02137	*
Customer.SegmentLoyalty Club Only	-1.499e+02	9.0122	-16.633129	< 2.2e-16	***
Customer.SegmentLoyalty Club and Credit Card	2.837e+02	11.9647	23.707798	< 2.2e-16	***
Customer.SegmentStore Mailing List	-2.453e+02	9.8382	-24.929693	< 2.2e-16	***
CityAurora	-1.885e+01	11.0982	-1.698837	0.08948	.
CityBoulder	3.954e+01	87.5614	0.451572	0.65162	
CityBrighton	8.845e+01	120.4283	0.734444	0.46275	
CityBroomfield	-4.848e-02	15.2218	-0.003185	0.99746	
CityCastle Pines	-6.497e+01	98.3595	-0.660511	0.50899	
CityCentennial	2.082e+00	18.9284	0.110006	0.91241	
CityCommerce City	-3.210e+01	44.4873	-0.721590	0.47062	
CityDenver	5.648e+01	27.3871	2.062130	0.03931	*
CityEdgewater	8.488e+01	47.4548	1.788750	0.07378	.
CityEnglewood	3.215e+01	23.9810	1.340677	0.18016	
CityGolden	9.256e+01	57.1187	1.620472	0.10527	
CityGreenwood Village	-2.273e+01	39.9056	-0.569506	0.56907	
CityHenderson	-1.154e+02	157.1146	-0.734334	0.46282	
CityHighlands Ranch	3.564e+00	33.4554	0.106541	0.91516	
CityLafayette	-4.286e+01	62.1797	-0.689359	0.49067	
CityLakewood	5.019e+01	28.8606	1.739207	0.08213	.
CityLittleton	1.478e+00	23.2569	0.063568	0.94932	
CityLone Tree	1.083e+02	138.4959	0.782201	0.43418	
CityLouisville	-2.295e+01	69.3343	-0.331046	0.74064	
CityMorrison	1.043e+02	75.5349	1.381322	0.16731	
CityNorthglenn	4.560e+01	39.9497	1.141544	0.25376	
CityParker	2.762e+01	31.9074	0.865663	0.38676	
CitySuperior	-4.787e+01	46.7553	-1.023859	0.30601	
CityThornton	9.658e+01	39.1145	2.469121	0.01362	*
CityWestminster	1.023e-01	17.5671	0.005825	0.99535	
CityWheat Ridge	2.314e+01	21.8514	1.058828	0.28979	
ZIP	-2.672e-01	0.1187	-2.251493	0.02445	*
Store.Number	-1.861e+00	1.1513	-1.616848	0.10605	
Avg.Num.Products.Purchased	6.714e+01	1.5262	43.995599	< 2.2e-16	***
X..Years.as.Customer	-2.374e+00	1.2314	-1.927967	0.05398	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.38 on 2341 degrees of freedom
Multiple R-squared: 0.8391, Adjusted R-Squared: 0.8368
F-statistic: 370 on 33 and 2341 DF, p-value: < 2.2e-16

Com base na exploração realizada acima, pudemos chegar ao modelo ideal, capaz de explicar 83,66% da variação da venda média e com variáveis dependentes de alta significância, utilizando as variáveis *Avg Num Products Purchased* e *Customer Segment*, contra 73,22% do modelo utilizando apenas a variável *Avg Num Products Purchased*.

$$Y = 303.46 + 66.98 * X_1 - 149.36 * X_2 + 281.84 * X_3 - 245.42 * X_4 + 0 * X_5$$

Onde:
Y = Avg.Sale.Amount
X_1 = Avg.Num.Products.Purchased
X_2 = Customer.Segment Loyalty Club Only
X_3 = Customer.Segment Loyalty Club and Credit Card
X_4 = Customer.Segment Mailing List
X_5 = Customer.Segment Credit Card

Report

Report for Linear Model Linear_Regression_37

Basic Summary

Call:
lm(formula = Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303.46	10.576	28.69	< 2.2e-16	***
Customer.SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer.SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer.SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16	***
Avg.Num.Products.Purchased	66.98	1.515	44.21	< 2.2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg.Sale.Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer.Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg.Num.Products.Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Residuals	44796869.07	2370			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Passo 3: Apresentação/Visualização

Aplicando o modelo obtido ao conjunto de dados de novos clientes, chegamos a uma receita esperada de **$ 47,224.87** (somatório da probabilidade de resposta positiva multiplicado pela venda média prevista) que, levados em consideração os custos de envio e a margem bruta, resultam num lucro de **$ 21,987.44**. Sendo assim, é recomendado o envio dos catálogos aos 250 clientes.