Business and Data Understanding

What decisions need to be made?

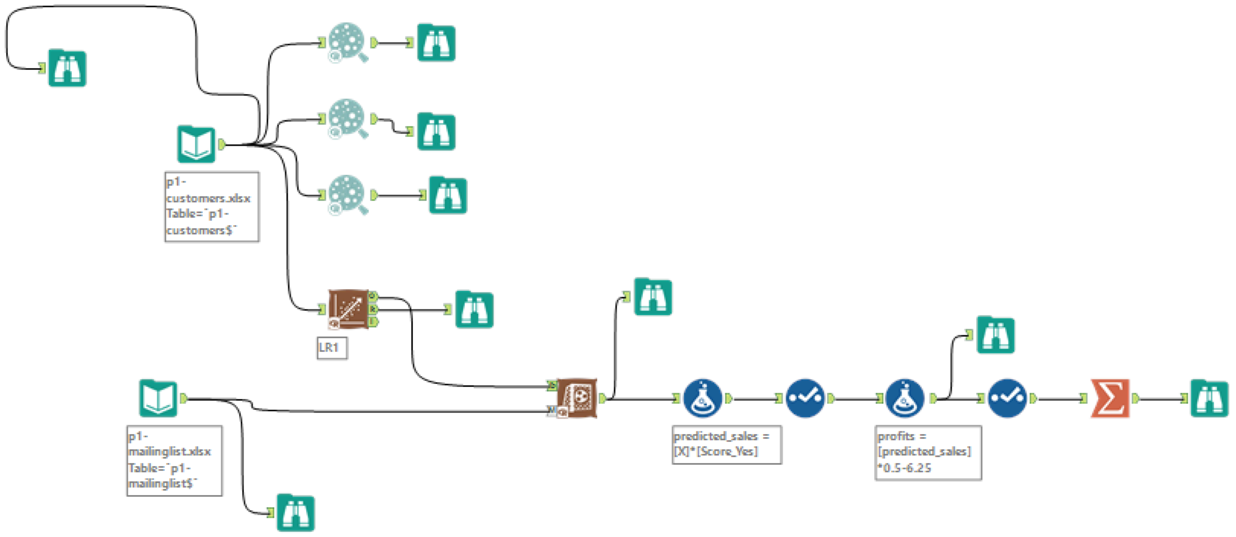
公司想依据现有的数据，预测一下如果给250位顾客派发新的产品册能否带来一定收入，从而决定是否进行派发。

Based on the existing data, the company would like to predict whether the distribution of new product catalog to 250 customers will bring certain profits, so as to decide whether to distribute them.

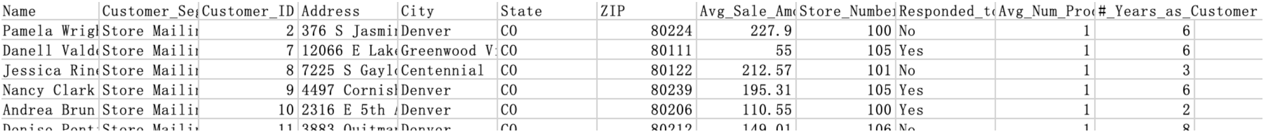
What data is needed to inform those decisions?

- 已有的顾客数据，来建立预测模型
- 通过线性回归模型，预测这250个顾客的可能花费，求出总和即可
- Existing customer data, to establish a prediction model
- Through the linear regression model, we can predict the possible costs of these 250 customers and calculate the sum

Analysis, Modeling, and Validation



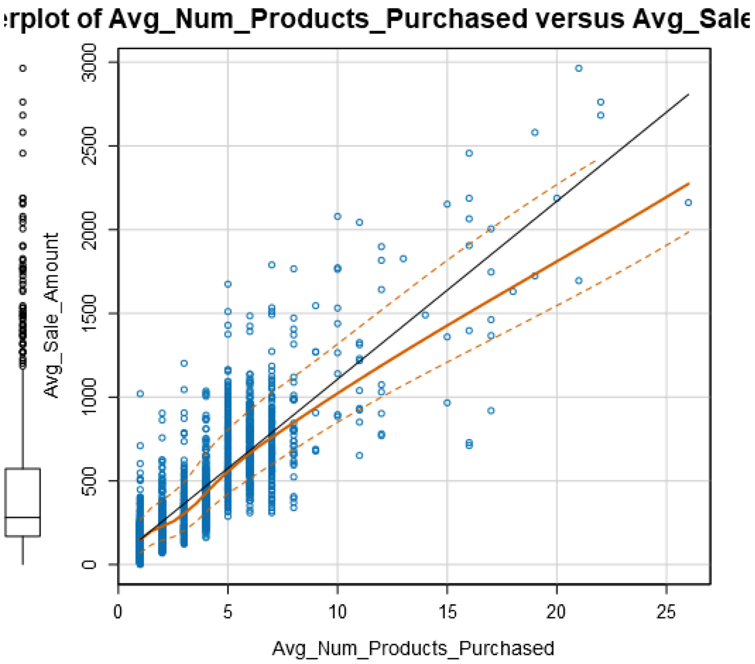
How and why did you select the predictor variables in your model?



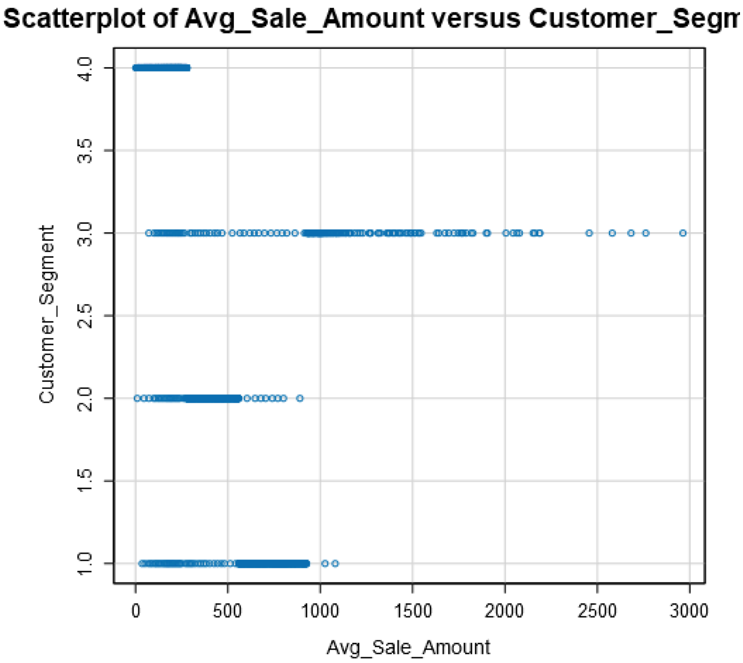
Name	Customer_Se	Customer_ID	Address	City	State	ZIP	Avg_Sale_Am	Store_Numbe	Responded_t	Avg_Num_Pro	#_Years_as_Customer
Pamela Wrigl	Store Mailir	2	376 S Jasmir	Denver	CO	80224	227.9	100	No	1	6
Danell Valde	Store Mailir	7	12066 E Lake	Greenwood V	CO	80111	55	105	Yes	1	6
Jessica Rine	Store Mailir	8	7225 S Gayl	Centennial	CO	80122	212.57	101	No	1	3
Nancy Clark	Store Mailir	9	4497 Cornisl	Denver	CO	80239	195.31	105	Yes	1	6
Andrea Brun	Store Mailir	10	2316 E 5th /	Denver	CO	80206	110.55	100	Yes	1	2
Denise Pont	Store Mailir	11	3883 Quitmar	Denver	CO	80212	149.01	106	No	1	8

The general content of the dataset used for training is shown in the figure above. Useless information can be ignored first, like *Name*,*ID*,*ZIP*,etc.

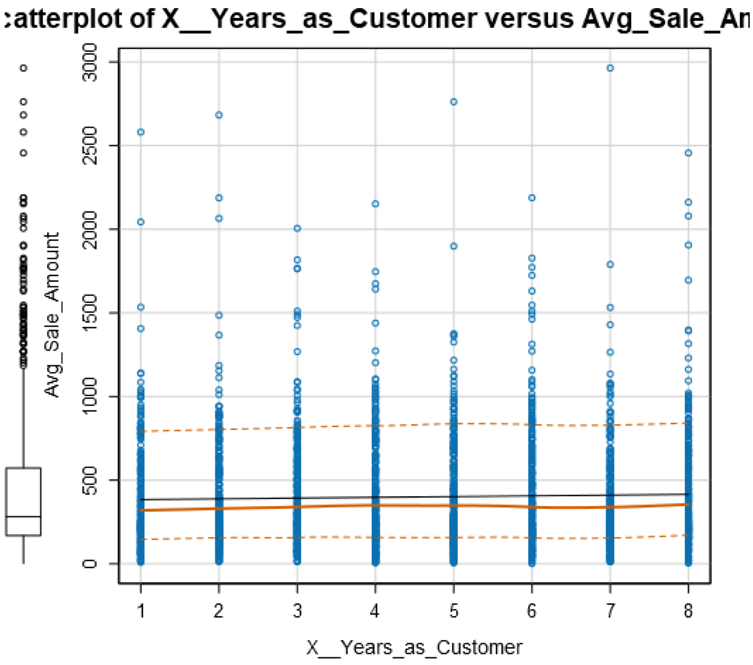
Because *Avg_sale_amount* is our forecast variable, we can explore the relationship between characteristic parameters that may be relevant.



ːrplot of Avg_Num_Products_Purchased versus Avg_Sale
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased

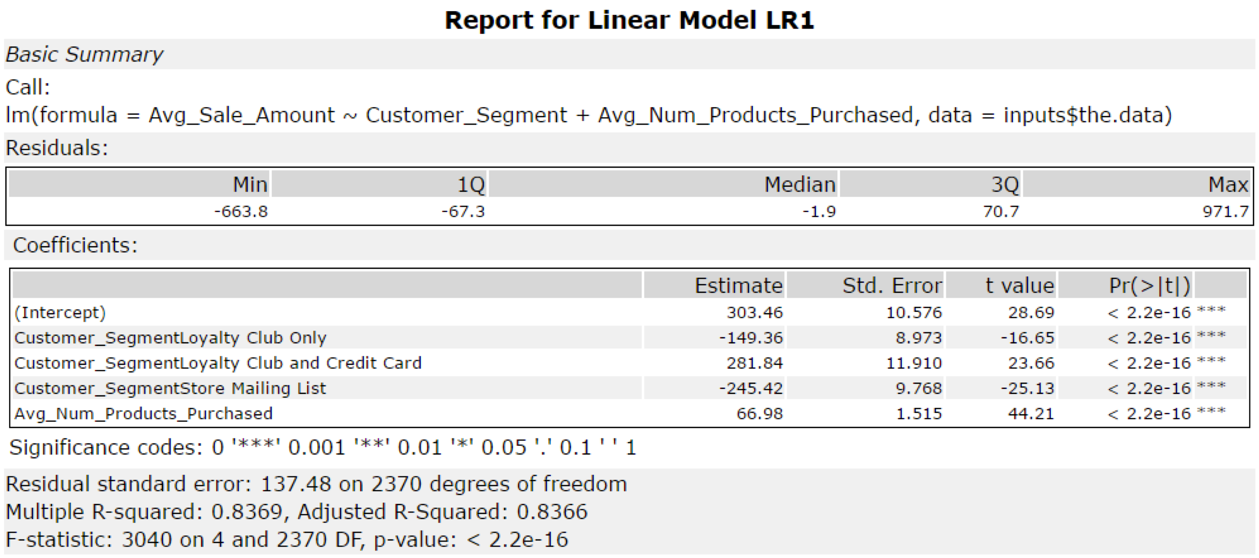


Scatterplot of Avg_Sale_Amount versus Customer_Segm
Customer_Segment
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount



From the scatter diagram of the above three possible variables and *Avg_sales_amount*, it's obvious that *Avg_num_products_purchased* has strong linear relationship with *Avg_sales_amount*,and different *Customer_segment* have different *Avg_sales_amount*. While *Years_as_customer*'s influence on *Avg_sales_amount* is not clear.(In fact, when using it as a parameter of linear regression, the corresponding p-value is relatively large)

Explain why you believe your linear model is a good model.



Report for Linear Model LR1

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

Report for this linear model is shown in figure above. For two variables we select, their p-value is far less than 0.05 which means their relationship with *Avg_sales_amount* is strong. R-squared and adjusted R-squared are also good enough which indicating the linear model we bulit is reliable.

Linear equation:

Avg_sales_amount=303.46-149.36(Customer_SegmentLoyalty Club Only)+281.34(Customer_SegmentLoyalty Club and Credit Card)-245.42(Customer_SegmentStore Mailing List)+66.98·Avg_Num_Products_Purchased

Conclusion

What is your recommendation? Should the company send the catalog to these 250 customers?

affirmative. The final calculated profit is **$22050**. The company is interested in it when projected profits is above $20000.