Project 1: Predicting Catalog Demand

The Business Problem

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The decision to be made is whether it is worth to send the company's catalog to these 250 new customers. The major indicator would be the expected profit contribution of this commercial action. If the expected profit exceeds USD 10,000, the catalog will be sent to these new customers. If it does not, the new customers will be removed from the catalog mailing list.

2. What data is needed to inform those decisions?

We need to know:

- ✓ The customer classification – For example,
  - whether the client is part of a loyalty program, mailing list or others
  - whether the client made a purchase after receiving the last catalog
  - the customer historical sales amount and number of purchase items
  - the customer geographic location and registered shopping area
  - the number of years as customer
  - the customer last purchase time and purchase frequency
  - the customer age, sex, marital status, family size and income
- ✓ The probability that these new customers will buy from the catalog,
- ✓ The expected revenue per new customer if the catalog is sent to them, and
- ✓ Calculate the expected profit contribution including the printing and distribution cost.

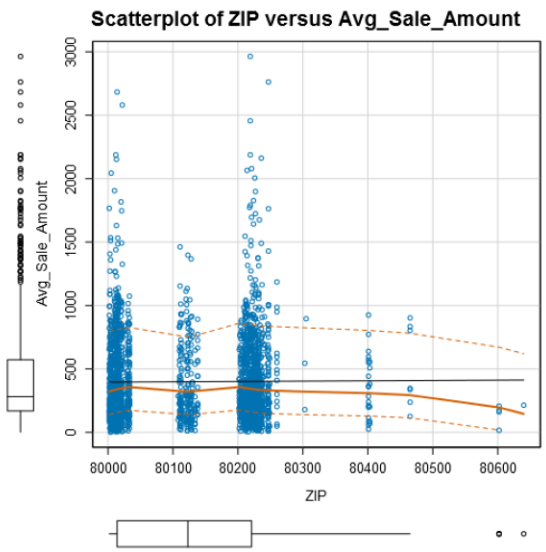
Step 2: Analysis, Modeling, and Validation

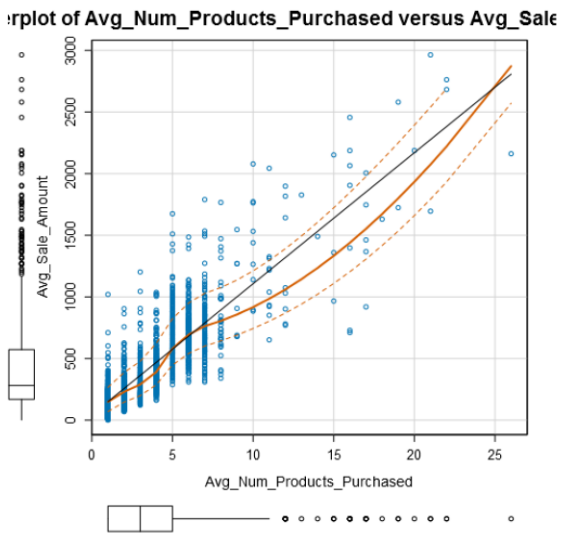
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

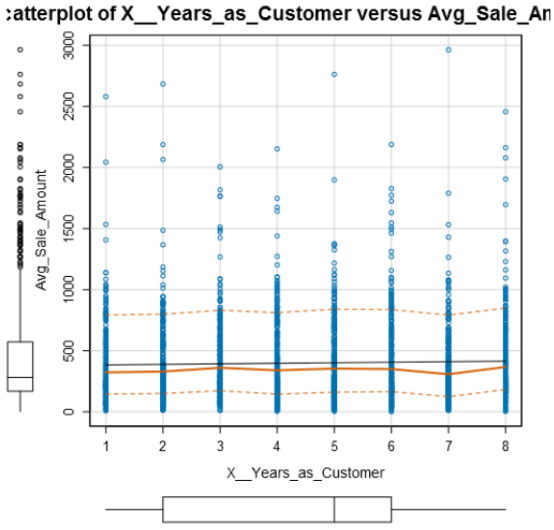
Important: Use the p1-customers.xlsx to train your linear model.

Based on the available data on the existing customers list, we have created a scatter-plot between the average sale amount (the target variable) and the different potential predictor variables. Since the customer segment and city are categorical variable, we run a regression model between the average sales amounts and customer segment and another model with the city.

We note from the charts and the model that only the average number of products purchased and the customer segment seems to have a linear relation with our target variable. The city is neither a statistically significant variable to predict the average sales.







Regression model testing the customer segment as a predictor variable of the avg sales amount

Record	Report
1	**Report for Linear Model reg**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment, data = the.data)
4	Residuals:

Min	1Q	Median	3Q	Max
-1002.00	-71.66	3.08	73.02	1889.00

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	682.7	8.354	81.72	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16	***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16	***

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

Residual standard error: 185.67 on 2371 degrees of freedom
Multiple R-squared: 0.7024, Adjusted R-Squared: 0.702
F-statistic: 1865 on 3 and 2371 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	192884931.52	3	1865.06	< 2.2e-16	***
Residuals	81736451.57	2371			

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

With a p-value < 0.05 and a R-squared, the customer segment seems to be a good predictor variable of the average sales amount.

Regression model testing the city as a predictor variable of the avg sales amount

Record | Report

1

Report for Linear Model reg

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ City, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Ma
-560	-232	-116	175	255

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	386.087	21.67	17.81399	< 2.2e-16	***
CityAurora	18.755	26.55	0.70630	0.48007	
CityBoulder	154.103	197.85	0.77890	0.43612	
CityBrighton	-291.157	241.83	-1.20398	0.22872	
CityBroomfield	7.409	37.39	0.19816	0.84294	
CityCastle Pines	-193.877	241.83	-0.80171	0.4228	
CityCentennial	-13.816	44.24	-0.31230	0.75484	
CityCommerce City	296.728	109.87	2.70065	0.00697	**
CityDenver	18.551	24.99	0.74237	0.45794	
CityEdgewater	76.875	100.69	0.76349	0.44525	
CityEnglewood	-9.806	50.41	-0.19450	0.8458	
CityGolden	-12.719	81.09	-0.15685	0.87538	
CityGreenwood Village	-60.038	93.58	-0.64157	0.52121	
CityHenderson	-171.697	341.31	-0.50305	0.61498	
CityHighlands Ranch	4.904	74.26	0.06604	0.94735	
CityLafayette	-41.955	153.86	-0.27267	0.78513	
CityLakewood	31.652	31.69	0.99872	0.31803	
CityLittleton	-9.727	45.62	-0.21322	0.83118	
CityLone Tree	468.783	341.31	1.37348	0.16973	
CityLouisville	-37.619	171.68	-0.21912	0.82658	
CityMorrison	126.608	130.55	0.96977	0.33226	
CityNorthglenn	-29.332	72.83	-0.40276	0.68716	
CityParker	-51.059	69.04	-0.73953	0.45966	
CitySuperior	-81.067	115.59	-0.70133	0.48317	
CityThornton	8.199	61.52	0.13327	0.89399	
CityWestminster	9.430	42.83	0.22016	0.82576	
CityWheat Ridge	43.875	51.17	0.85744	0.39129	

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

Residual standard error: 340.62 on 2348 degrees of freedom
Multiple R-squared: 0.008008, Adjusted R-Squared: -0.002976
F-statistic: 0.7291 on 26 and 2348 DF, p-value: 0.8374

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
City	2199299.15	26	0.73	0.83744
Residuals	272422083.94	2348		

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

With most of p-values > 0.05 and a R-squared of 0.008008 , the city is not a significant predictor variable of the average sales amount.

Based on this analysis, we run a regression model using the average sales amount as target variable and the average number of products purchased as predictor variables. We obtained the following equation and report.

Avg Sales Amount

$$= 303.46 - (149.36 \text{ x Loyalty Club only}) + (281 \text{ x Loyalty Club and credit card}) - (245.42 \text{ x Store mailing list}) + (0 \text{ x Credit card only}) + (66.98 \text{ x Avg Number of Products Purchased})$$

Record	Report

Report for Linear Model reg

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = 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

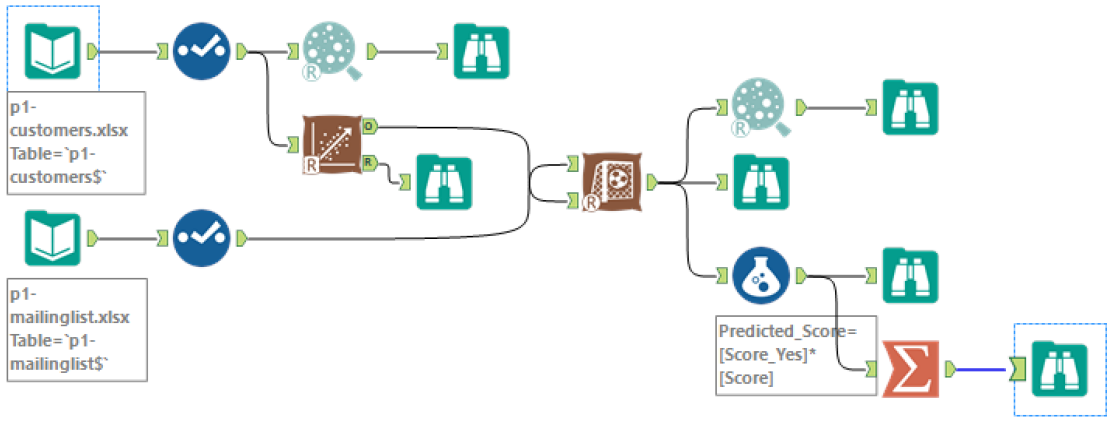
The model R-squared is high at 0.8369. The p-values of our coefficients are well below 0.05 and are statistically significant.

Based on the available data, this is the best model we can have since the other potential predictors variables do not appear to be statistically significant for the prediction.

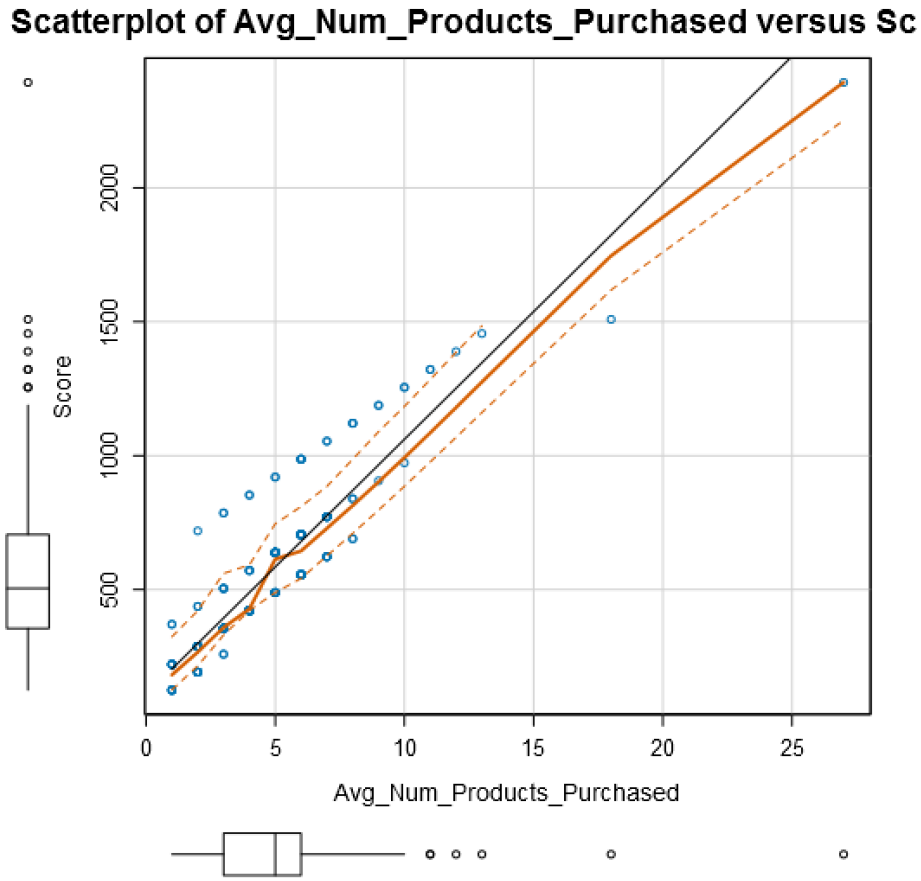
Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

We applied the model to the batch of 250 new customers using the score tool in Alteryx.



We created a scatter-plot of the score against the average number of products purchased to visualize the results.



Then we use the formula tool to multiply the predicted score with the probability of the new customer to buy the catalog. Finally, we use the summarize tool to add the results and we obtain a predicted sales of USD 47,224.87.

To determine the expected profit contribution,

- we calculate the average gross profit margin: USD 47,224.87 * 50% = USD 23,612.44 and
- we deduct the printing and distribution cost from the profit margin

$$\textbf{Expected profit contribution} = \text{USD } 23{,}612.44 - (250 \text{ x USD } 6.5) = \textbf{USD } \mathbf{21{,}987.44}$$

In conclusion, according to the model, if the catalog is sent to the 250 new customers, the expected profit contribution is USD 21,987.44, which is higher than the minimum fixed by the manager. Therefore, the company should send the catalog to these 250 new customers.