Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

1. What decisions needs to be made?

We want to predict how much profit the company can expect from 250 new customers if they send out the catalog.

2. What data is needed to inform those decisions?

We need to know the customer segment and the average number of products he purchased before, these to data will help us in our dissection.

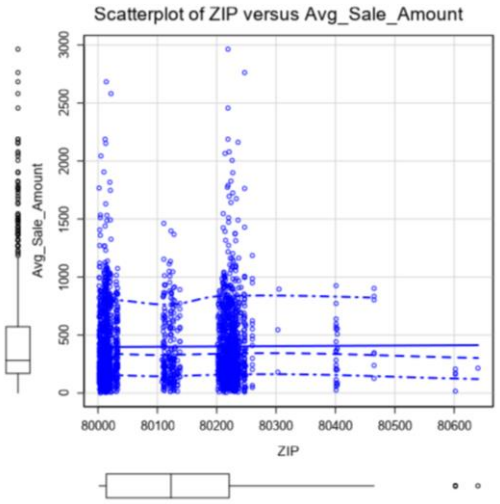
Step 2: Analysis, Modeling, and Validation

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

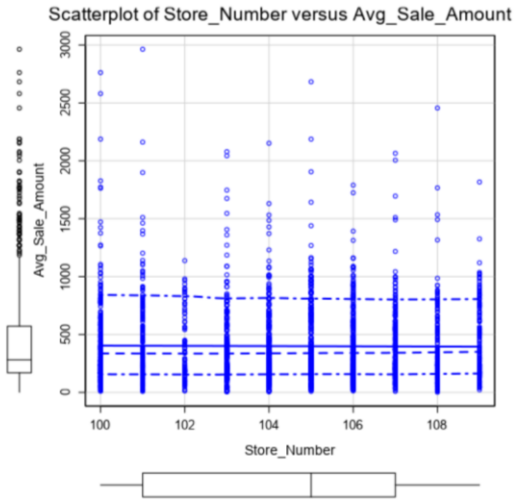
At the begging I searched for the target to exclude it, so depend in our decision we want to predict which is profit the company expect for each customer the target will be (Avg_Sale_Amount).
After that I draw the scatter plot for each numeric feature to see if there is a liner relation between each feature and our target.

1- Customer ID Vs Avg_Sale_Amount

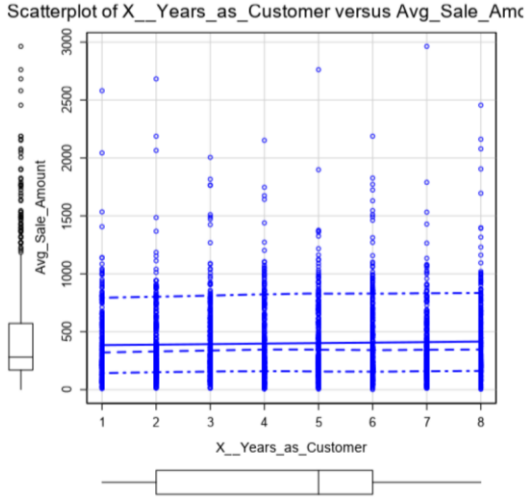
2- Zip Code Vs Avg_Sale_Amount



3- Store Number Vs Avg_Sale_Amount



4- Years As customer Vs Avg_Sale_Amount



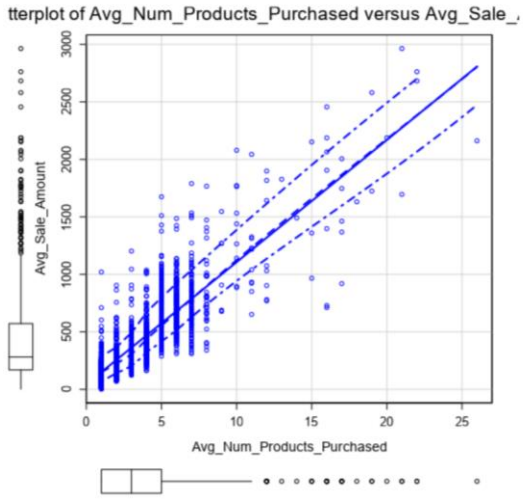
For the categorical feature I run the linear regression model and check each one if these features less significant to be eliminate.

CityAurora	-6.0853	14.49	-0.419833	0.67465
CityBoulder	126.9674	107.95	1.176189	0.23964
CityBrighton	-62.9990	131.98	-0.477348	0.63316
CityBroomfield	-9.2150	20.40	-0.451708	0.65152
CityCastle Pines	-85.4308	131.96	-0.647414	0.51743
CityCentennial	-27.0212	24.14	-1.119260	0.26314
CityCommerce City	43.5787	60.04	0.725774	0.46805
CityDenver	10.5272	13.64	0.771979	0.4402
CityEdgewater	37.0387	54.93	0.674262	0.50021
CityEnglewood	2.3193	27.50	0.084328	0.9328
CityGolden	1.2895	44.26	0.029135	0.97676
CityGreenwood Village	47.8220	51.10	0.935885	0.34943
CityHenderson	-469.9963	186.34	-2.522240	0.01173 *
CityHighlands Ranch	-63.6199	40.52	-1.569954	0.11656
CityLafayette	33.0268	83.94	0.393441	0.69403
CityLakewood	-14.2164	17.31	-0.821314	0.41155
CityLittleton	-16.9903	24.89	-0.682661	0.49489
CityLone Tree	170.4837	186.34	0.914902	0.36034
CityLouisville	-0.9317	93.66	-0.009947	0.99206
CityMorrison	-36.2462	71.26	-0.508630	0.61106
CityNorthglenn	14.9483	39.75	0.376062	0.7069
CityParker	6.9027	37.68	0.183202	0.85466
CitySuperior	-28.3948	63.10	-0.449993	0.65276
CityThornton	25.2262	33.57	0.751502	0.45243
CityWestminster	-8.3505	23.38	-0.357147	0.72101
CityWheat Ridge	-3.4180	27.94	-0.122356	0.90263

So, the result was I eliminate all the features except two which are (number of products purchased and customer segment)

To double check if I used the right features or not, I draw the scatter plot to see the linear relation between the feature and the target

1- If the number of products purchased increase the profit will be increase at an approximate linear fashion.



2- For the categorical features (customer segment) I used the significant to see if there is a relation with the target or not.

	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

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

After running the model, I came up with this result:

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 degrees of freedom (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 adjusted R-squared value is 0.836. Whenever the adjusted R-squared greater than 0.5 that means high explanatory power of the model.

For the variable which I selected I came up with this result:

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

Depend on the P-Value all the variable selected are most significant which means there is actual relationship between these variables and our target.

3. What is the best linear regression equation based on the available data?

Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased -149.36 (If Type: Club Only) + 281.84 (If Type: Club and Credit Card) – 245.42 (If Type: Mailing List)

If the customer segment type Credit Card Only we must include the 0 coefficient for all other types.

Step 3: Presentation/Visualization

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

To decide if we should send a catalog for these 250 customers or not, we need to find the total profit and it's should be between 20,000$ and 25,000$ as the information provided.
So, to find the total profit of these customers who will buy if we send a catalog for them

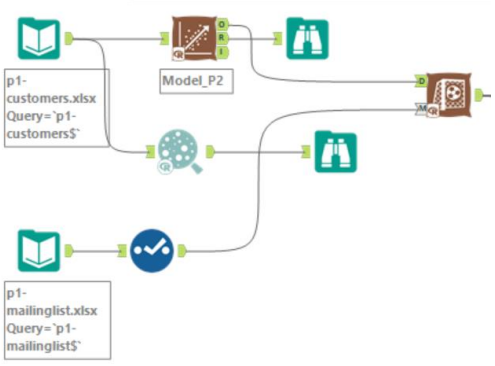
First, I build the linear regression model to predict the Avg_Sale_Amount for each customer, then I multiplied each Avg_Sale_Amount with the possibility of customer who said will buy when we send a catalog for him, After that I get the sum of this then multiplied it with gross margin to get the real cost of all product sold in catalog after that I subtracted out the cost of sending the catalog for each customer finally I multiplied the result with 250 which is the number of customer then I came up with the total profit.

After this calculation I came up with total profit 21,987.43$ which is more than 20,000. So, my recommendation will be sending the catalog for the customers.

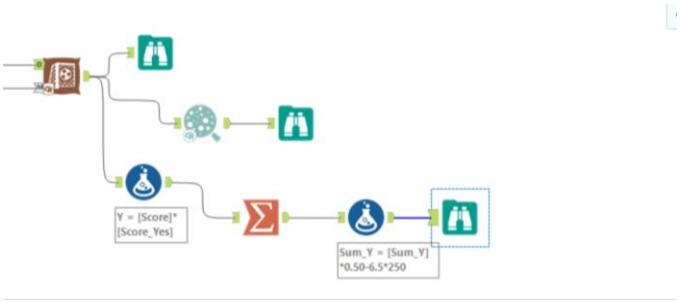
2. **How did you come up with your recommendation?**

As I said in the above question, we need to find the total profit and to do that I follow multiple steps I will explain it in detail now:

1. First build the machine learning model using training dataset then predict the Avg_Sale_Amount.



2. Then multiply the predicted value with Score_Yes which is the possibility of the customer will buy when sending the catalog. After that we get the sum of these.
3. Then we need to multiply the summation with gross margin which is 50%.
4. After that we need to subtract the 6.5$ from the profit which is the cost of each catalog.
5. The last step we need to multiply this with 250 which is the number of the new customers the manager wants to send the catalog for them.



Note: Score represent predicted Avg_Sale_Amount.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

After all I did the calculation and as I explained in above question by using Alteryx software the profit will be 21,987.43$