Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   - The management needs to decide whether or not to send out catalogs to a list of 250 new customers. This decision will be based on the condition that the expected profit that will be generated from sending out these catalogs will exceed $10,000. If this condition is not met, the catalogs will not be sent.

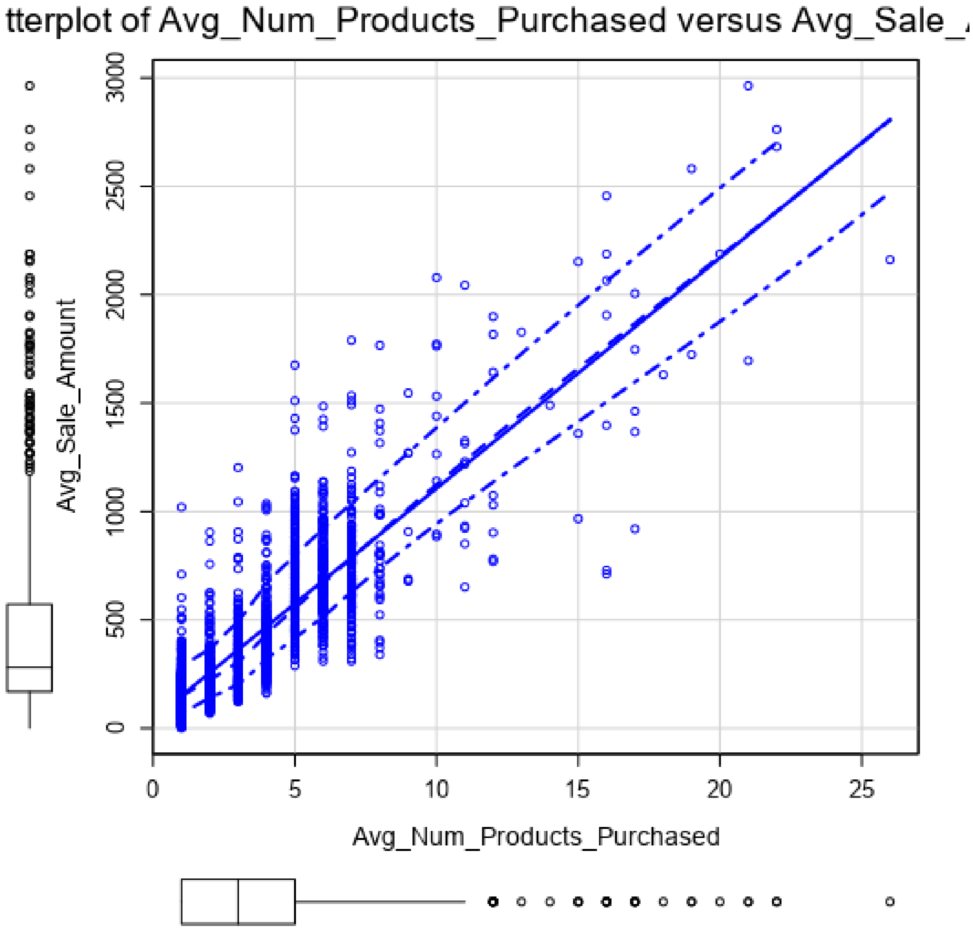
2. What data is needed to inform those decisions?

   - In order to assist the management make a decision, a dataset containing information about sales in the past will be needed. This will be helpful to be able to carry out proper predictive analysis to help inform the decision of the management.

Step 2: Analysis, Modeling, and Validation

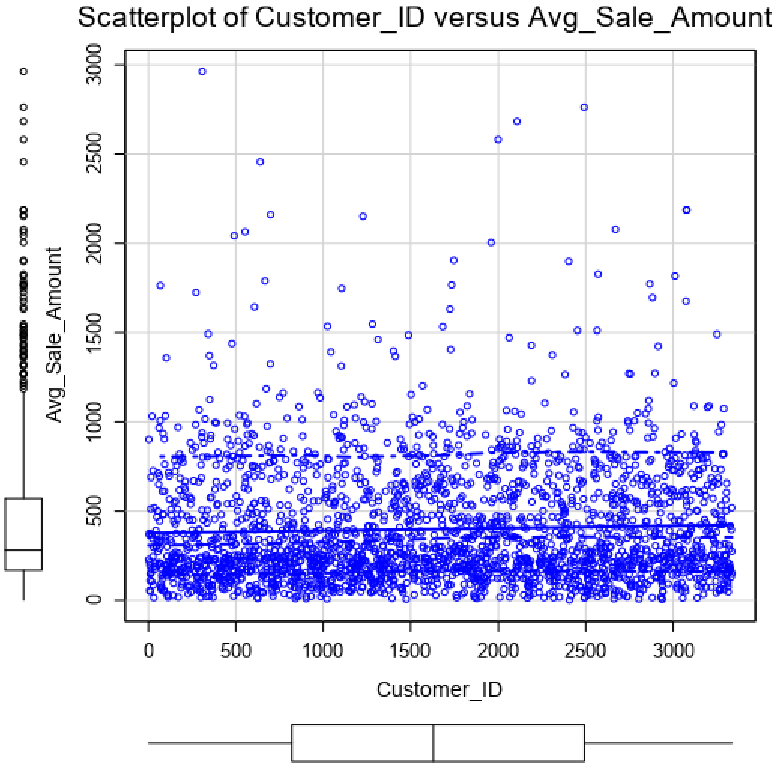
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

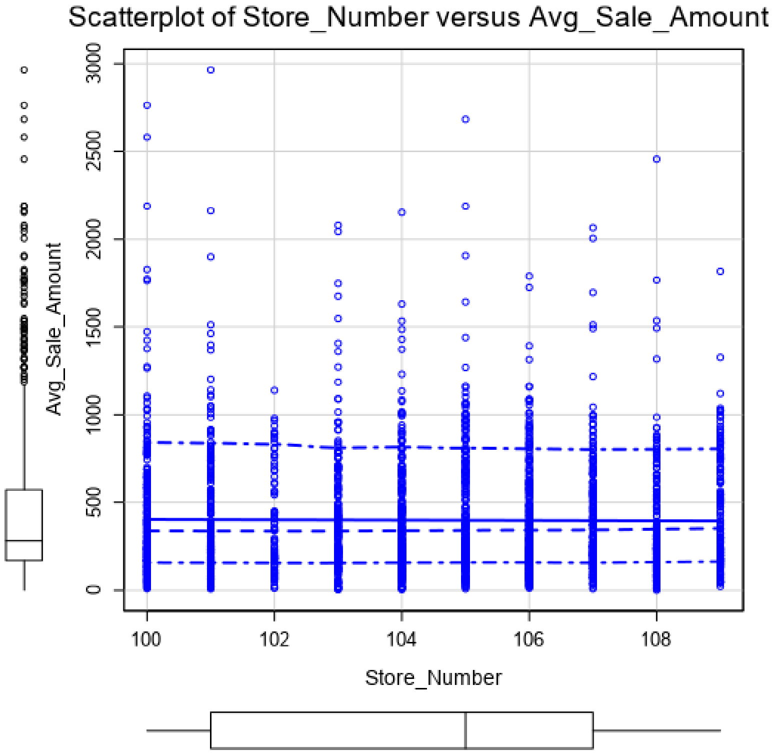
   - I have decided to choose two predictor variables (Customer Segment and Avg_Num_Products_Purchased) to be used in my linear regression model.
   - I chose to use the 'Customer Segment' as a predictor variable because it is a categorical variable which dummy variables will be assigned to and it has the capability of improving the strength of my model.
   - I also chose to use the 'Avg_Num_Products_Purchased' as a predictor variable because it is the only numeric predictor variable that displays a high level of correlation with the target variable (Avg_Sale_Amount). See below the scatter plot showing the relationship between 'Average Number of Products' and 'Avg_Sale_Amount':

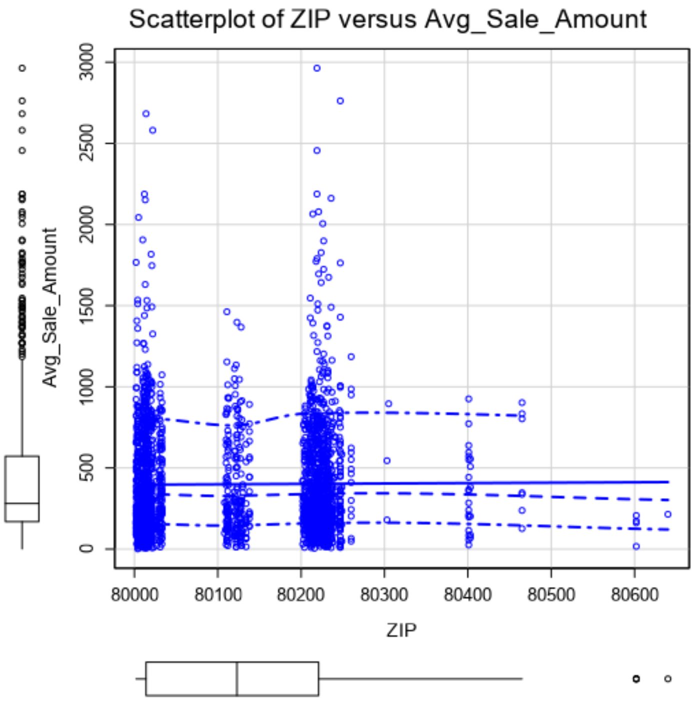


The plot generally shows that the higher the number of products purchased, the higher the sales amount.

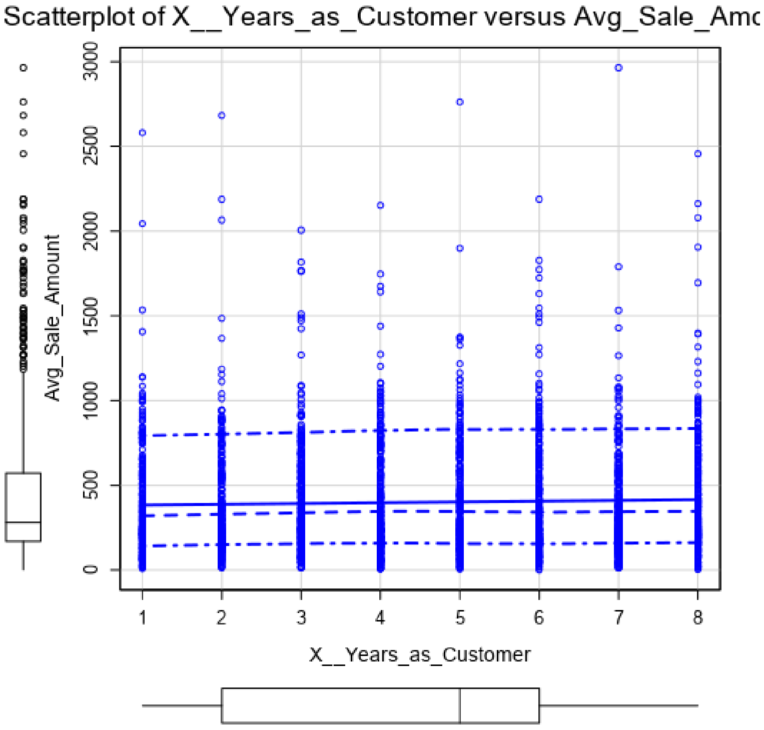
- Other numeric predictor variables such as ‘Customer ID’, ‘Store Number’, ‘Zip’ and ‘X_Years as Customer’ were excluded from the model because they do not display a linear relationship with the target variable. See below the scatter plots showing their relationships with the target variable respectively:


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


Scatterplot of Store_Number versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
Store_Number
100
102
104
106
108



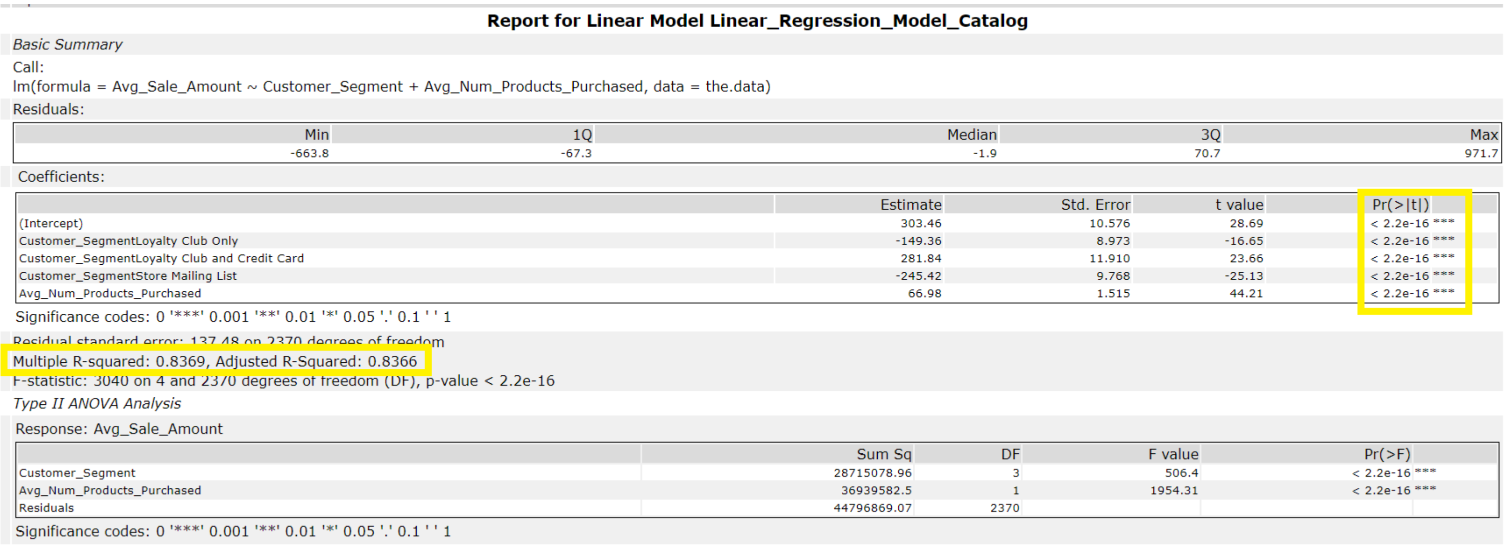
Scatterplot of ZIP versus Avg_Sale_Amount



Scatterplot of X__Years_as_Customer versus Avg_Sale_Amc

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

    - See below the report of my model using ‘Customer Segment’ and ‘Avg_Num_Products_Purchased’ as predictor variables:



Report for Linear Model Linear_Regression_Model_Catalog

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 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

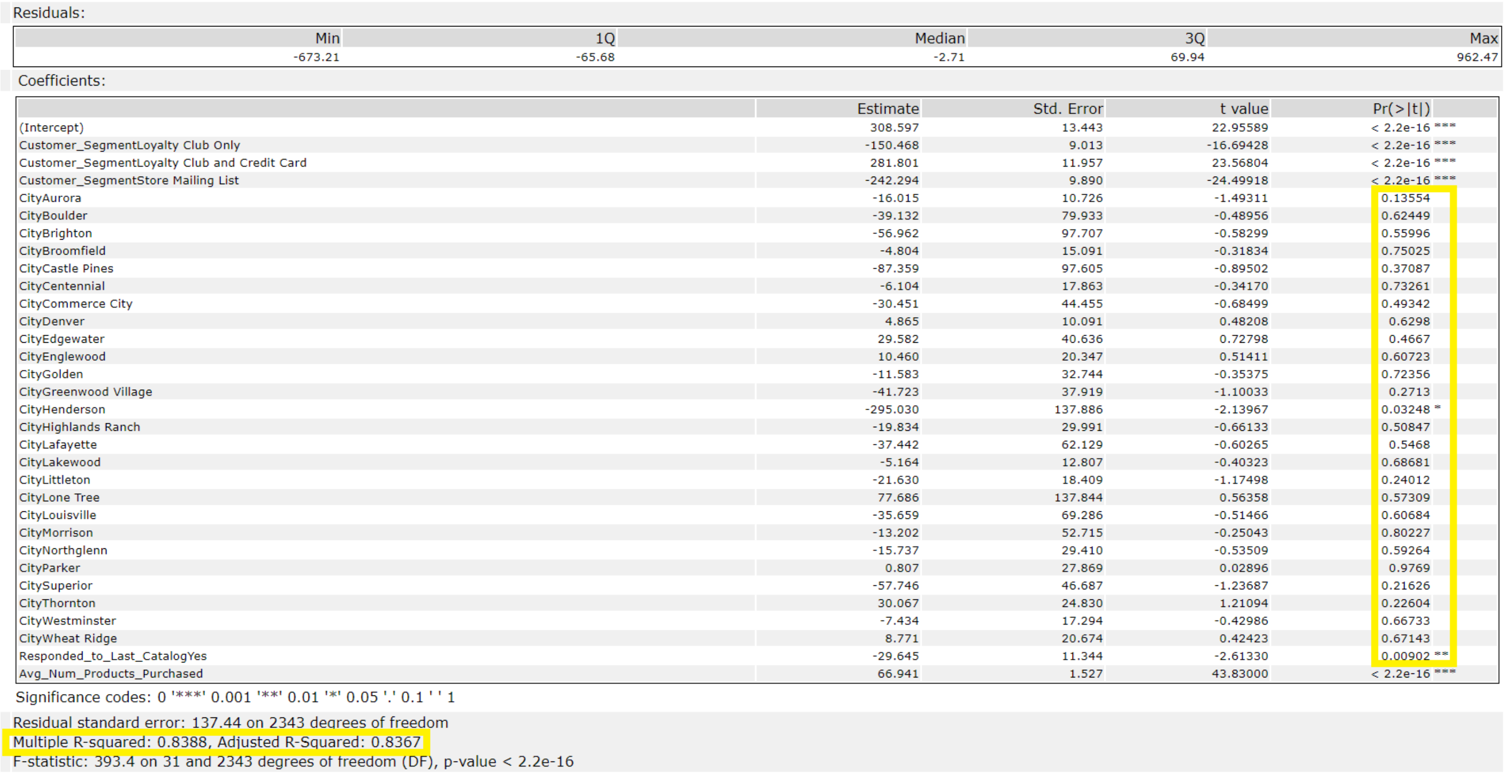
Response: Avg_Sale_Amount

	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

When a linear model is been created, it is considered as a good model fit for use if the P-value <= 0.05 and the R-squared value > 0.7. Considering the result obtained from my regression model, the P-values for each of the predictor variables suggests them to be highly statistically significant and the R-squared value is also high which makes the model fit for use.

However, I tried adding the other categorical variables such as ‘City’ and ‘Responded_to_Last_Catalog’, the R-squared value seem to be good but they were not as statistically significant as that of ‘Customer Segment’, so I decided not to consider them in my model. See below the report from the model:



Residuals:

Min	1Q	Median	3Q	Max
-673.21	-65.68	-2.71	69.94	962.47

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	308.597	13.443	22.95589	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.468	9.013	-16.69428	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.801	11.957	23.56804	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.294	9.890	-24.49918	< 2.2e-16	***
CityAurora	-16.015	10.726	-1.49311	0.13554	
CityBoulder	-39.132	79.933	-0.48956	0.62449	
CityBrighton	-56.962	97.707	-0.58299	0.55996	
CityBroomfield	-4.804	15.091	-0.31834	0.75025	
CityCastle Pines	-87.359	97.605	-0.89502	0.37087	
CityCentennial	-6.104	17.863	-0.34170	0.73261	
CityCommerce City	-30.451	44.455	-0.68499	0.49342	
CityDenver	4.865	10.091	0.48208	0.6298	
CityEdgewater	29.582	40.636	0.72798	0.4667	
CityEnglewood	10.460	20.347	0.51411	0.60723	
CityGolden	-11.583	32.744	-0.35375	0.72356	
CityGreenwood Village	-41.723	37.919	-1.10033	0.2713	
CityHenderson	-295.030	137.886	-2.13967	0.03248	*
CityHighlands Ranch	-19.834	29.991	-0.66133	0.50847	
CityLafayette	-37.442	62.129	-0.60265	0.5468	
CityLakewood	-5.164	12.807	-0.40323	0.68681	
CityLittleton	-21.630	18.409	-1.17498	0.24012	
CityLone Tree	77.686	137.844	0.56358	0.57309	
CityLouisville	-35.659	69.286	-0.51466	0.60684	
CityMorrison	-13.202	52.715	-0.25043	0.80227	
CityNorthglenn	-15.737	29.410	-0.53509	0.59264	
CityParker	0.807	27.869	0.02896	0.9769	
CitySuperior	-57.746	46.687	-1.23687	0.21626	
CityThornton	30.067	24.830	1.21094	0.22604	
CityWestminster	-7.434	17.294	-0.42986	0.66733	
CityWheat Ridge	8.771	20.674	0.42423	0.67143	
Responded_to_Last_CatalogYes	-29.645	11.344	-2.61330	0.00902	**
Avg_Num_Products_Purchased	66.941	1.527	43.83000	< 2.2e-16	***

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

Residual standard error: 137.44 on 2343 degrees of freedom
Multiple R-squared: 0.8388, Adjusted R-Squared: 0.8367
F-statistic: 393.4 on 31 and 2343 degrees of freedom (DF), p-value < 2.2e-16

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

- The best linear regression eqation is as stated below;

  Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 * (Customer Segment: Loyalty Club Only) + 281.84 * (Customer Segment: Loyalty Club and Credit Card) – 245.42 * (Customer Segment: Store Mailing List) + 0 * (Customer Segment: Credit Card Only)

Step 3: Presentation/Visualization

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

    - Based on the result obtained from my analysis, my recommendation to the management is to proceed to send out the catalog to the 250 new customers.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

    - I built a model using the p1-customers dataset and then applied the model to the p1-mailinglist dataset using the score tool to obtain the predicted sale amount.
    - Since the Score_Yes indicates the probability that a customer will make purchase after receiving the catalog, I multiplied each predicted sale amount by their corresponding Score_Yes values to obtain the new predicted sale amount.
    - From the project details, the average gross margin (price - cost) on all products sold through the catalog is 50%, so I multiplied each new predicted price by 0.5 to obtain the gross profit.
    - Also since the costs of printing and distributing for each catalog is $6.50, I subtracted 6.5 from each gross profit to obtain the net profit.
    - I then made a summation of all the net profit to arrive at the total expected profit if the catalog was sent to the entire 250 new customers.

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

    - The expected profit, if the catalog is sent to the new 250 customers, is $21,987.44.