Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**

Should the company send out print catalog to 250 new customers from their mailing list in the coming months?

2. **What data is needed to inform those decisions?**

p1-customers.xlsx - This dataset includes the following information on about 2,300 customers and will be used to build model.

p1-mailinglist.xlsx - This dataset is the 250 customers that need sales prediction. This is the list of customers that the company would send a catalog to. Use this dataset to estimate how much revenue the company can expect if they send out the catalog.

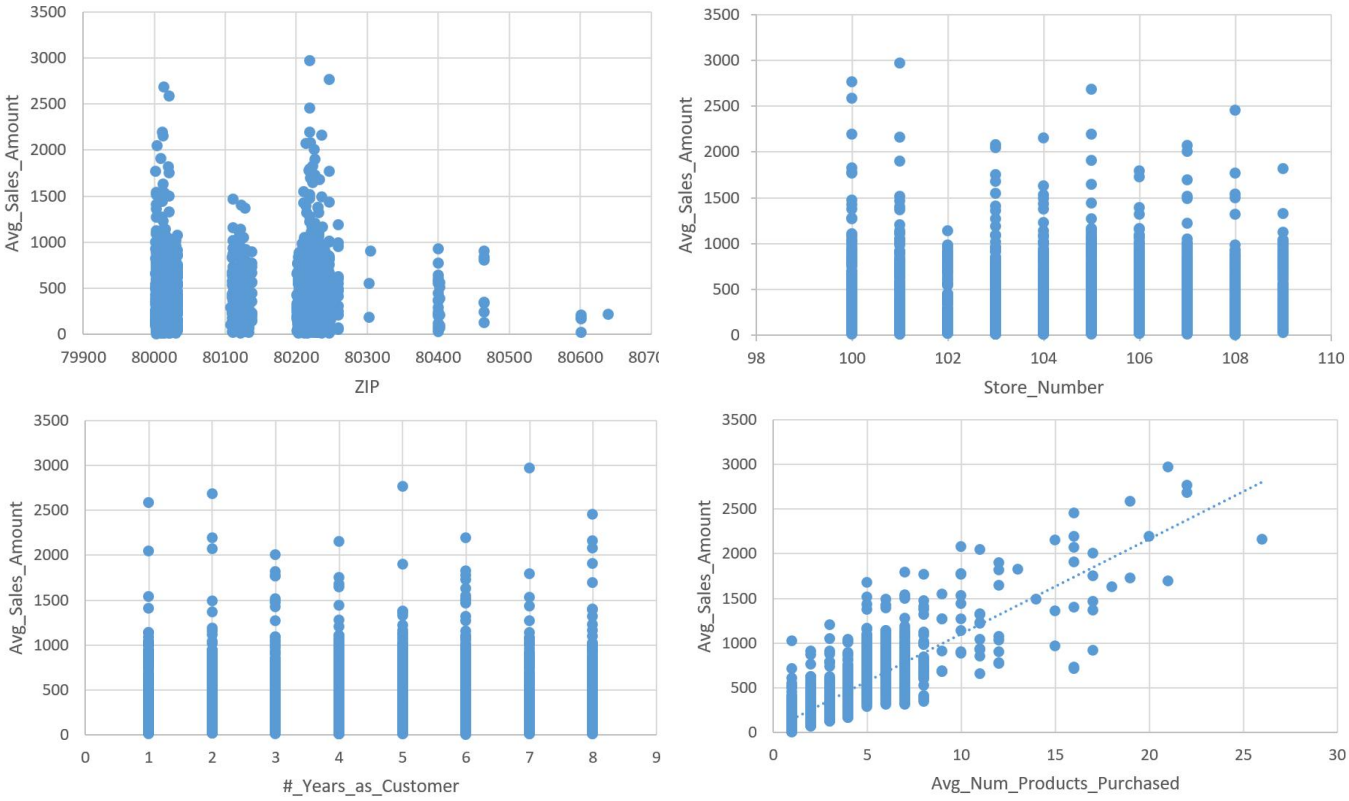
Step 2: Analysis, Modeling, and Validation

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.**

Firstly there are 5 independent variables taken into consideration for the linear model:

- Customer_Segment
- ZIP
- Store_Number
- Avg_Num_Products_Purchased
- #_Years_as_Customer

Secondly, using scatter plot to test the relationship between each variable and the dependent variable (Avg_Sales_amount) as the following:



From the scatter plots above, we can find that ZIP, Store_Number and #_Year_as_Customer do not have linear relationship with the dependent variable. Only Avg_Num_Products_Purchased has a positive linear relationship with the dependent variable.

Then create 3 dummy variables for the column Customer_Segment. Now we have 5 dependent variables to build the linear regression model. Obtaining the result as below:

Multiple R	0.914947784							
R Square	0.837129447							
Adjusted R Square	0.836785694							
Standard Error	137.4060735							
Observations	2375							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	229893646.7	45978729	2435.25871	0			
Residual	2369	44727736.4	18880.43					
Total	2374	274621383.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	313.7602446	11.8606814	26.45381	2.468E-135	290.501853	337.018636	290.5018532	337.018636
Loyalty Club and Credit Card	282.6215498	11.91020278	23.72937	7.486E-112	259.266049	305.977051	259.2660487	305.977051
Loyalty Club Only	-149.1066517	8.968665228	-16.6253	8.6059E-59	-166.6939	-131.51941	-166.6938981	-131.519405
Store Mailing List	-245.4831405	9.76235525	-25.1459	7.058E-124	-264.62679	-226.3395	-264.6267859	-226.339495
Avg_Num_Products_Purchased	67.01829126	1.514350075	44.25548	0	64.0487024	69.9878801	64.04870245	69.9878801
#_Years_as_Customer	-2.340072599	1.222907498	-1.91353	0.05580039	-4.7381525	0.05800726	-4.738152462	0.05800726

We can find that only #_Years_as_Customers has p_value > 0.05 so it's not statistically significant in the model. Leaving out this variable to run the model again.

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.**

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.91481							
R Square	0.836878							
Adjusted R Square	0.836602							
Standard Error	137.4832							
Observations	2375							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	4	229824514	57456129	3039.744	0			
Residual	2370	44796869.07	18901.63					
Total	2374	274621383.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	303.4635	10.57571483	28.69437	1.1E-155	282.72486	324.20208	282.72486	324.2020827
Loyalty Club and Credit Card	281.8388	11.90985741	23.66433	2.6E-111	258.483946	305.19358	258.4839461	305.1935838
Loyalty Club Only	-149.356	8.972754792	-16.6455	6.35E-59	-166.95098	-131.7605	-166.950984	-131.7604598
Store Mailing List	-245.418	9.767775616	-25.1252	1.1E-123	-264.57201	-226.2635	-264.572015	-226.263474
Avg_Num_Products_Purchased	66.9762	1.515040358	44.20754	0	64.0052631	69.947147	64.00526313	69.94714671

I believe my linear model is a good model for several reasons:

- Adjusted R Squared is approximately 0.84 which means the model can explain well at 84% in the variation of the dependent variable.

- All independent variables selected for the model are statistically significant as all p_value < 0.05.

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

Y = 303.46

+ 281.84 (If Customer Segment: Loyalty Club and Credit Card)

- 149.36*(If Customer Segment: Loyalty Club Only)

- 245.42*(If Customer Segment: Store Mailing List)

- 0*(If Customer Segment: Credit Card Only)

+ 66.98*Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

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

The company should send the catalog to these 250 customers as the expected profit exceeds $10,000.

2. How did you come up with your recommendation?

I come up with my recommendation based on the predicted profit if sending out the catalog to 250 customers. The process is as follows:

- Select appropriate continuous variables and create dummy variables. Then training the linear regression model in excel with training data from 2,300 customers in the file p1-customers.xlsx.

- Get the best linear regression equation and then plug in to the file p1-mailinglist.xlsx to get sales prediction for each of 250 customers.

- Calculate expected revenue for each customer by multiply the probability that a person will buy our catalog with sales prediction.

- Calculate expected profit for each customer by multiply the expected revenue with gross margin (50%) then subtract out the $6.50 (the costs of printing and distributing per catalog).

- Calculate final expected profit by summing all expected profit for each customer. This number is used to justify whether or not the company should send out the catalog.

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

The expected profit is $21,987.96.