Project 1: Predicting Diamond Prices

Step 1: Business and Data Understanding

A company that manufactures and sells high-end home goods is going to send this year`s catalog to 250 new customers from their mailing list. They would like to estimate how much profit the company can make from the new catalog. The calculated profit should be over $10,000 otherwise the management doesn`t prefer to make the effort for the new catalog. There are some conditions listed below that should be included to the estimation but not limited to.

- This calculation will be conducted based on the previous catalog data and results.
- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- The estimated revenue should be multiple by the gross margin first before you subtract out the $6.50 cost when calculating your profit.
- The new client data contains the information below:
    - Score_No: The probability that the customer WILL NOT respond to the catalog and not make a purchase.
    - Score_Yes: The probability that the customer WILL respond to the catalog and make a purchase.

.

Key Decisions:

1. What decisions needs to be made?

The decision is whether the company should send the new catalog to the new 250 costumers to make over $10,000 profit or not.

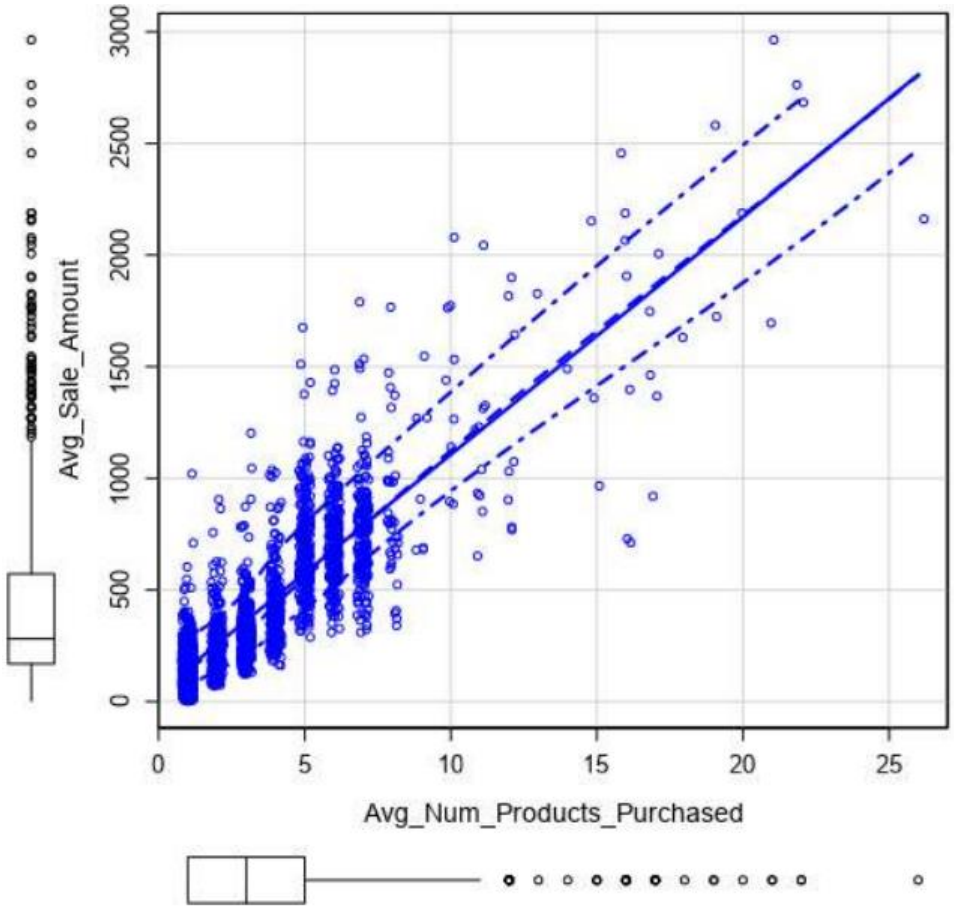
2. What data is needed to inform those decisions?

- p1-customers.xlsx - This dataset includes the rich information about 2,300 customers.
- p1-mailinglist.xlsx - This dataset is the 250 new customers that will be used to predict the sales
- The cost of each catalog is $6.50.
- Gross margin of sales for catalog is (%50)
- The probability (Score_Yes) that the new costumers will make the purchase

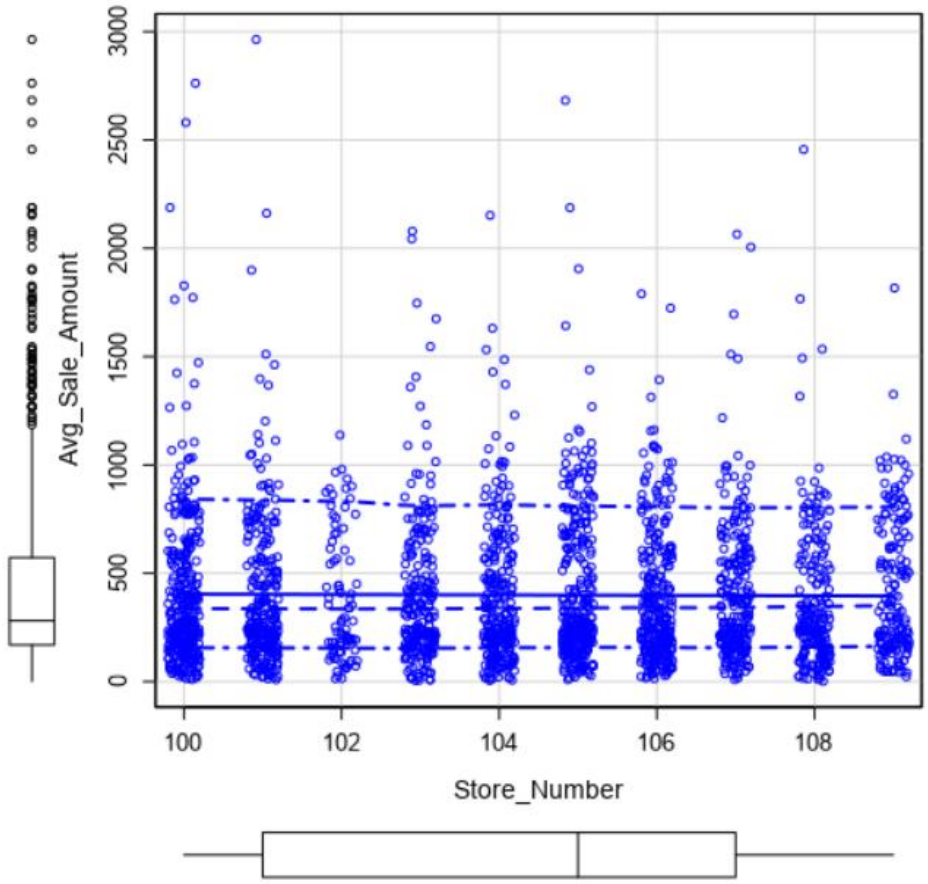
Step 2: Analysis, Modeling, and Validation

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.

In order to select the correct predictor in the model, scatter plots were created to understand the relationships and multiple reports with different predictors were generated using linear regression tool. As an example, the scatter plot below shows the correlation between average sale amount and average number of products purchased. It can be seen that these two datasets have a strong relationship.



The scatter plot below shows the relationship between average sales amount and store numbers. Even though there is a relationship, it is not as clear as the previous example. This study was continued in the linear regression tool to select the best predictors for our model.



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.

The result of P-values and R square values show that the final linear model is a good model. As shown in the report below, P-values are less than ‘0.5’ and very close to ‘0’. Additionally, R squared values (multiple and adjusted) are above ‘0.7’. These findings indicate that there is a high relationship between target and predictors.

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

As a comparison, multiple linear models were tested based on the outputs of scatter plots. The example below shows that the correlation of store_number is less significant than others. Therefore, it was removed from the model.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	962.703	140.23	6.865	8.46e-12	***
Customer_SegmentLoyalty Club Only	-287.675	11.35	-25.350	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	390.446	15.69	24.884	< 2.2e-16	***
Customer_SegmentStore Mailing List	-519.697	10.18	-51.039	< 2.2e-16	***
Store_Number	-2.671	1.34	-1.993	0.04637	*
Responded_to_Last_CatalogYes	-50.705	15.17	-3.343	0.00084	***

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)

Y = 303.46 + (
– 149.36 * <Customer_Segment = "Loyalty Club Only">
+ 281.84 * <Customer_Segment = "Loyalty Club and Credit Card"
- 245.42 * <Customer_Segment = "Store Mailing List>
+ 66.98 * <Avg_Num_Products_Purchased>
- 0.00 * <Customer_Segment = "Cash">
)

Step 3: Presentation/Visualization

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

Based on the findings of the model, it is my recommendation that the company should send the catalog to the 250 new costumers since the profit exceeds $10,000.

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

Firstly, the provided datasets were explored and studied. Some issues that may affect the analysis were found. These issues were cleaned from the datasets using “Select” tool and “Data Cleansing” tool. After the data normalization, scatterplot was utilized to have a better understanding of the relationships. Based on the findings and provided limitations/rules, multiple models were designed and tested. Finally, the linear regression model with the lowest p-value and highest R squared value was selected and applied to the new dataset using “Score” tool. From the result of the “score” tool, expected profit was calculated using the Formula below.

```
catalog_cost = 6.5
revenue_yes = predicted_value * score_yes
each_expected_profit = (revenue_yes * 0.5) – catalog_cost
expected_profit = sum(each_expected_profit)
```

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

The expected profit from the new catalog is $21,987.44.