Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The main decision that needs to be made is whether to send the catalog to 250 new customers or not.

Key Decisions:

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

We need to decide whether we should send product catalog to 250 new customers or not.

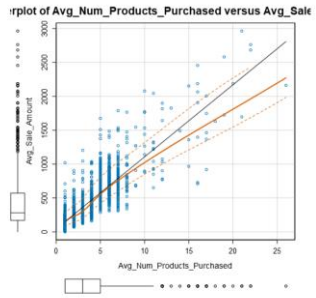
2. **What data is needed to inform those decisions?**

To make that decision we need to calculate the total profit that will be generated if we send the catalogs to 250 new customers. To find out the total profit we need to first calculate the total revenue. We have two data sets that will help us. The first one, p1_customers, contains data about old customers, including customer segment, average number of products purchased, number of years as a customer, etc. More importantly, in this file we have customers' response to previous catalog request and average sale amount they made. Using all available data about old customers we can choose our target variable as "average sale amount" and make a regression model using appropriate predictor variables to calculating total sale amount. We can test a few numeric or categorical variables (for example: "average number of products purchased", "number of years as a customer", "customer segment", "city", etc.) and find out the most strongly correlated ones to our target variable. After we build a model, we need to use it to predict the sale amount for the new 250 customers. The data about these customers is in the second file: "p1_mailinglist". In this file we have almost all the data as in "p1_customers" except "average sale amount" which is logical, as we need to calculate that value through predicting model, and also we don't have customers' response to catalog which we don't know yet for new customers. However, we have two other data columns "score_yes" and "score_no" which shows the probability of whether a customer will or will not buy a catalog. After we predict the sale amount for each customer using our model, we need to multiply it with "score_yes", the probability that a customer will buy the catalog, in order our predicted sale amount will be closer to reality. To calculate the profit after we predicted the revenue we need data about the costs. We can find it in project details: profit margin is 50% so we need to multiply predicted revenue with this number, and after that need to subtract total costs of printing and distributing catalogs, which is $6.5 per each costumer, so total cost is $6.5x250. After we subtracted the costs we can have the final profit amount, which is necessary to make the decision.

Step 2: Analysis, Modeling, and Validation

1. **How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.**

As a target variable I chose “Average sale amount” as sale amount is the first indicator of income. As predictor variable I choose one numerical and one categorical variable. The numerical variable is “Average number of products purchased”. As it is visible in below scatterplot there is a positive relationship between this variable and targets variable.



The categorical variable is “Customer segment”, as it has low p-value when doing regression analysis.
Other variables including “# of years as a customer” are not statistically significant and have high P-value. For example, in case of “# of years as a customer” p-value is 0.055 which is too high to use this variable in a good 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 linear model I built is a good one, because there is 0 (***) statistical significance between the target variable and chosen predictor variables. The p-values for all the predictor variables are “<2.2e-16, which is a very good indicator, and R-Squared value is 0.8369.

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

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 linear regression equation is the following:
Average sales amount = 303.46 - (0 * Credit card only) + (149.36 * Loyalty Club Only) + (281.84 * Loyalty Club and Credit Card) - (245.42 * Store Mailing List) + (66.98 * Average Number of Products Purchased)

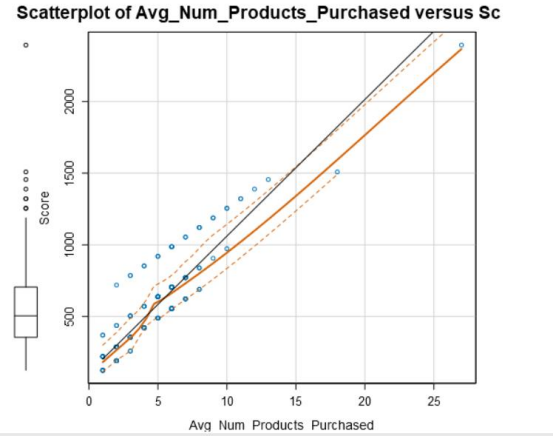
Step 3: Presentation/Visualization

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 because the predicted profit from this action will be $21978 which exceeds $10000, the minimum amount for the company to be profitable.

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

I used linear regression tool in Alteryx to build a linear model. Afterwards, I used score tool to calculate the average amount of sales per each customer for the new 250 customers.
Below is the scatterplot representing the relationship between the score and avg num products purchased. As it is visible, there is strong correlation which proves we have selected a right predictor variable for our target variable.



Then I used formula tool (score*score yes) to calculate the sales amount per each customer taking into account the probability that a particular customer will buy the catalog. I named this variable as “Income”. After this step I summarized the incomes of all the customers to one total income using summarize tool. As mentioned in the project details, the average gross margin is 50%. So, in the final formula I counted 50% of income and reduced total costs **($6.5 * 250)** from this amount.

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

The expected profit amount is **$21987.43**.