Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Answer: The questions have to be answered or the decision has to be made whether it is worthwhile to send the catalog to the 250 customers. This should only be done if the expected profit exceeds $10,000. The historical data from the Excel "p1-customers" form the basis for the linear regression model. The equation created in this way is the foundation for further evaluation. Fortunately, we can draw on sufficient data here (Data rich)

2. What data is needed to inform those decisions?

Answer: In principle, two data sets are required. On the one hand, the historical data set "p1-Customers", which looks at the purchase relationship of 2375 customers and provides information on whether the mailing of the catalog has led to an increase in sales. On the other hand, the "mailing list" data set looks at the total of 250 potential customers who are sent catalogs and their possible purchase intention. For the linear regression model, the data "Customer_Segment" and "Avg_Num_Products_Purchased" are used. The "Sale" was selected as the target varibale. To create a link between the Excel file "Customer" and "Mailing", the variables in the "Select" area were renamed and adjusted accordingly. Other important factors are the varibals "yes_score", "margin" and "catalog costs". The variable "Yes_score" is supplied by the Excel file "p1-mailinglist" and is an important part of the further calculation. The average gross margin (price - cost) on all products sold through the catalog is 50%. Where the cost of printing and distributing the catalog is $6.50 per catalog.

Step 2: Analysis, Modeling, and Validation

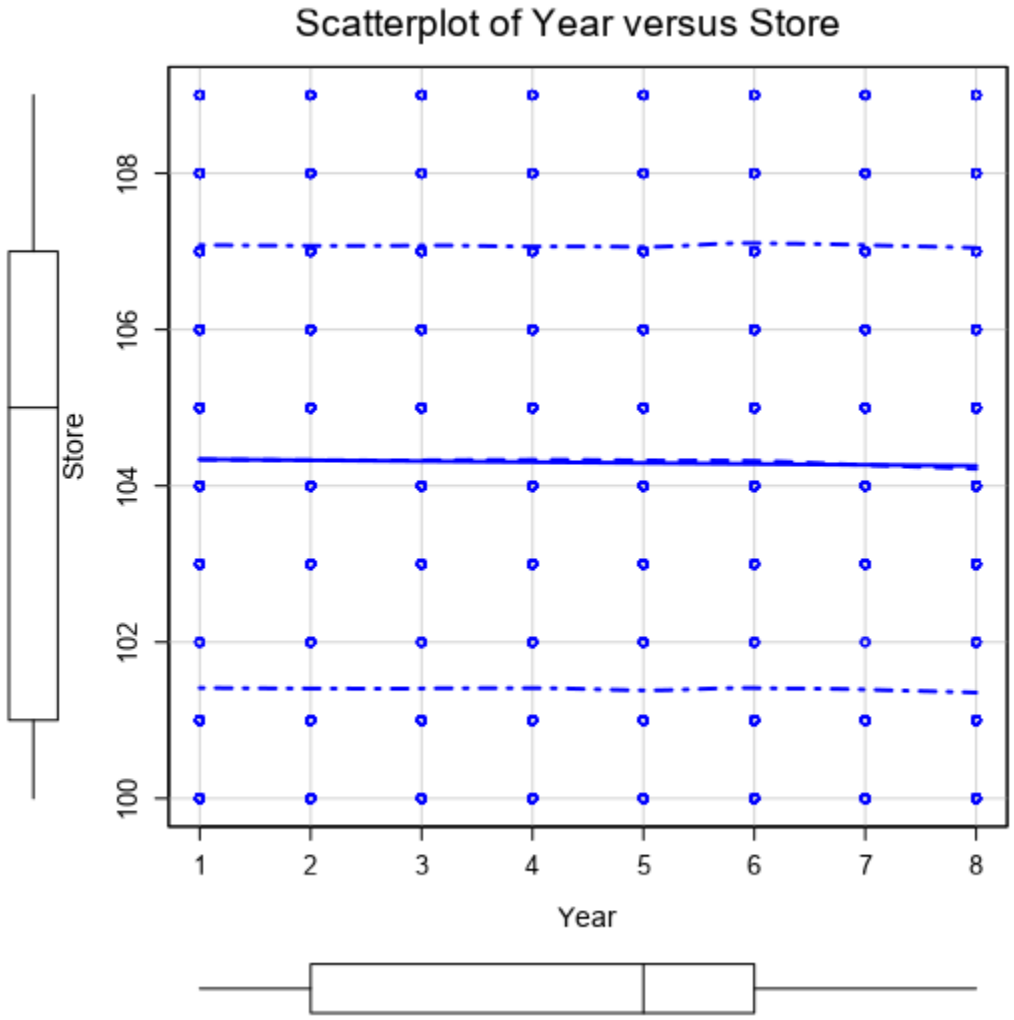
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

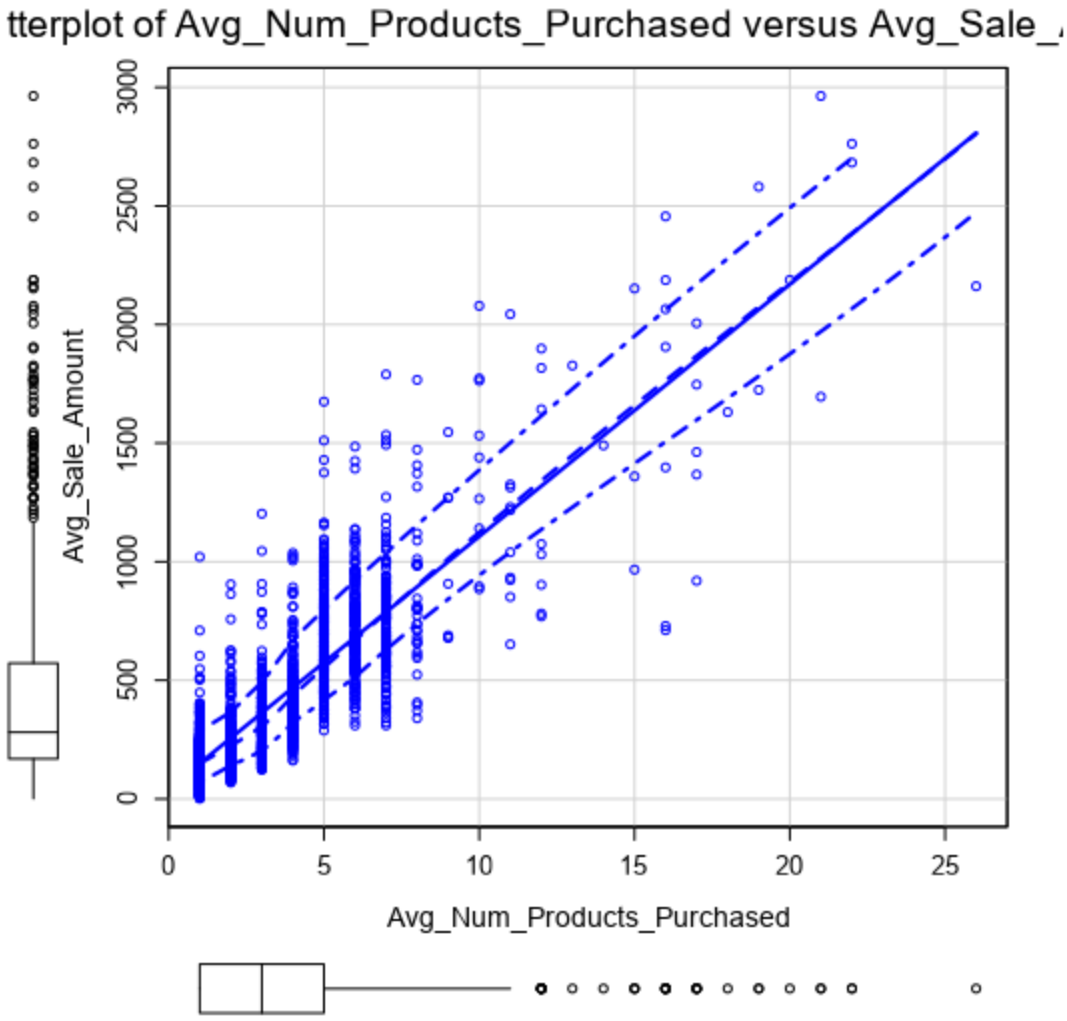
At the minimum, answer these questions:

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.

Answer: The best way to answer this question is by the use of a scatterplot and a linear regression by Alteryx. The Scatterplot “Years_as_Customer” with “Store_Number” shows no correlation and should be an example, which I can’t use for my further analysis.



The opposite is the following scatterplot. With this two numerical variables “Avg_Sale_Amount” and “Avg_Num_Products_Purchased” I created a Scatterplot, which will be the foundation of my further analysis.



In the table below you can see that the P-Value for the variable “Customer_Segemnt” and “Avg_Num_Products_purachsed is very low and so meets the goal to have a p-value less than 0.05. Also, the three stars beside the variables show you that you can use these variables for an analysis. That is a good indicator for using these variables and to us them for my analysis.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958
CustomerLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16 ***
CustomerLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16 ***
CustomerStore Mailing List	-245.8485	9.770e+00	-25.1625	< 2.2e-16 ***
ZIP	0.0225	2.659e-02	0.8460	0.39761
Store	-1.0002	1.006e+00	-0.9939	0.32037
Product	66.9646	1.515e+00	44.1928	< 2.2e-16 ***
Year	-2.3528	1.223e+00	-1.9239	0.05449 .

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.

Answer:
P-Value: Statistical significance is indicated by a p-value. If the p-value is less than 0.05, the result of a statistical test is called "significant. In this case, all p-values are 0.

R-squared -Value: The R-squared -Value is a key figure, which can assume values between 0 and 1. The closer the value is to 1, the more accurate the prediction.

If the value is 0, then no variance of the dependent variable can be explained by the independent variables. The value 1 again describes the case that the variance of the criterion is completely explainable by the predictors. In our case, the R-value is 0.7 and thus a good value.

Besides the P-Value the R-Value/adjusted R-value is very important. As you can see in this case the R-value is also high so it is another indicator for me to continue with this model.

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

Answer: Y = 303.46 – 149.36 * (CustomerLoyalty_Club_only) + 281.84 * (CustomerLoyalty_Club_and_Credit_Card) - 245.52 * (CostumerStore_mailing_List) + 66.98 * (Avg_Num_products_purchased) + 0 * (Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

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

Answer: If the company mails the total of 250 catalogs, a profit of $21.987can be expected. Since the premise was at least $10,000, the recommendation is to mail the catalog. The recommendation is Yes, the company should send out the catalogs.

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

Answer: The recommendation is based on historically provided data. Based on this data it is possible to make a forecast regarding the 250 potential customers or to make a decision whether it makes sense to send the catalogs. The model is based on linear regression as an aid.

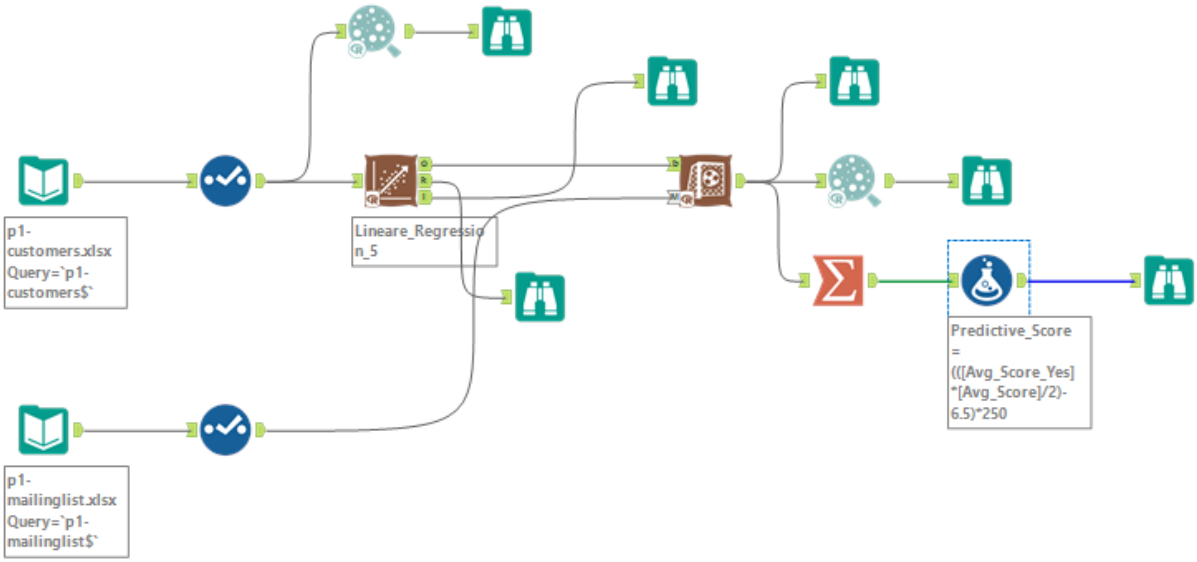
The following formula was used over the total 250 potential customers and resulted in $21.987.

(Average Score*Score_Yes*0.5) - (6.5*250) = $21.987

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

Answer: The expected profit is $21.987.

Alteryx Workflow:


p1-
customers.xlsx
Query=`p1-
customers$`
Lineare_Regressio
n_5
p1-
mailinglist.xlsx
Query=`p1-
mailinglist$`
Predictive_Score
=
(([Avg_Score_Yes]
*[Avg_Score]/2)-
6.5)*250