Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

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

The company wants to send out a catalog to some new customers, but they only want to send them if the profit contribution from them is greater than $10,000. For this to be done it means the company will predict the expected profit from these new customers.

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

In order to inform those decisions, we have been provided with two datasets, p1-customers.xlxs and p1-mailinglist.xlxs. This means the Company already has some information for more than 2,000 customers in the first dataset. This information is the Sales made by the Company for the already existing Customers and can be considered as numeric data, therefore Linear Regression Model can be used to predict the profit from these new customers in the p1-mailinglist dataset. Customers dataset was used to build or train the model while mailing dataset was used to make predictions and validate the model.

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 words 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?

When I ran the model with almost all the predictor variables, I realized some of them except the Customer_segment and Avg_Num_Products_Purchased had p-values greater than 0.05, this means there was no actual relationship between these predictor variables and the observation. Then I dropped insignificant variables that have less predictive power and used Customer_segment and Avg_Num_Products_Purchased that have a very high predictability for the target variable. Below are visualizations showing the degree of correlation between the predictor variables and the target variable.

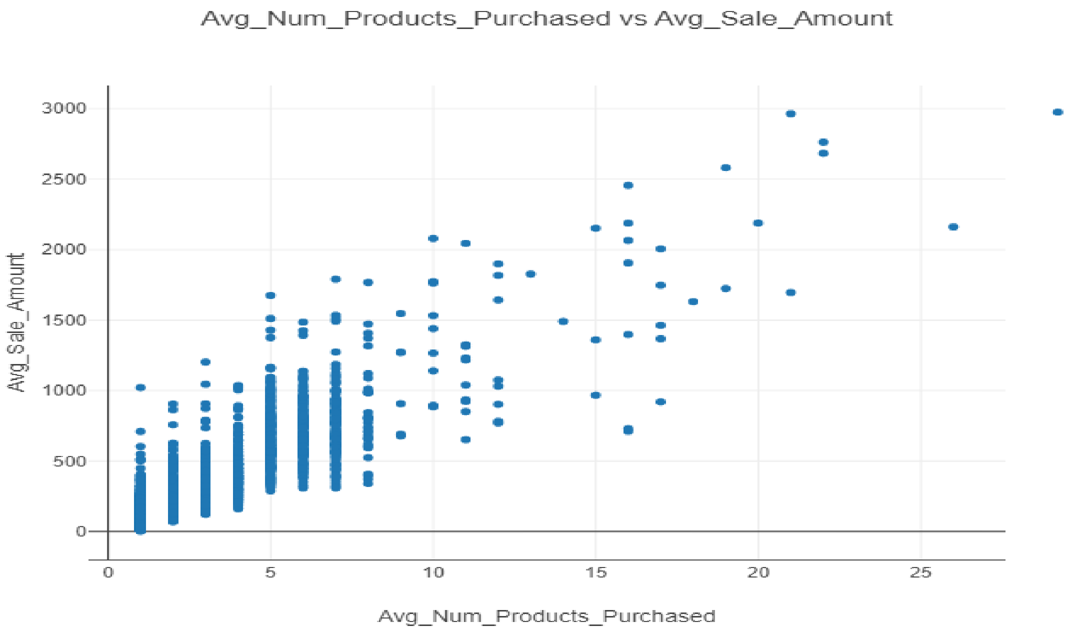


Fig.1

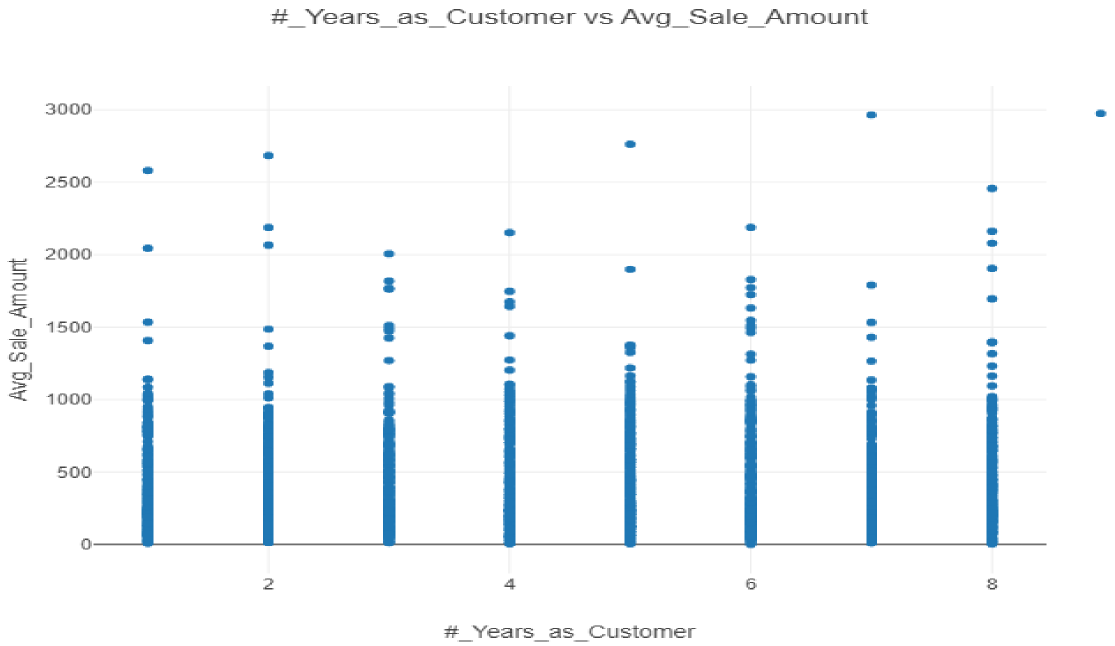


Fig. 2

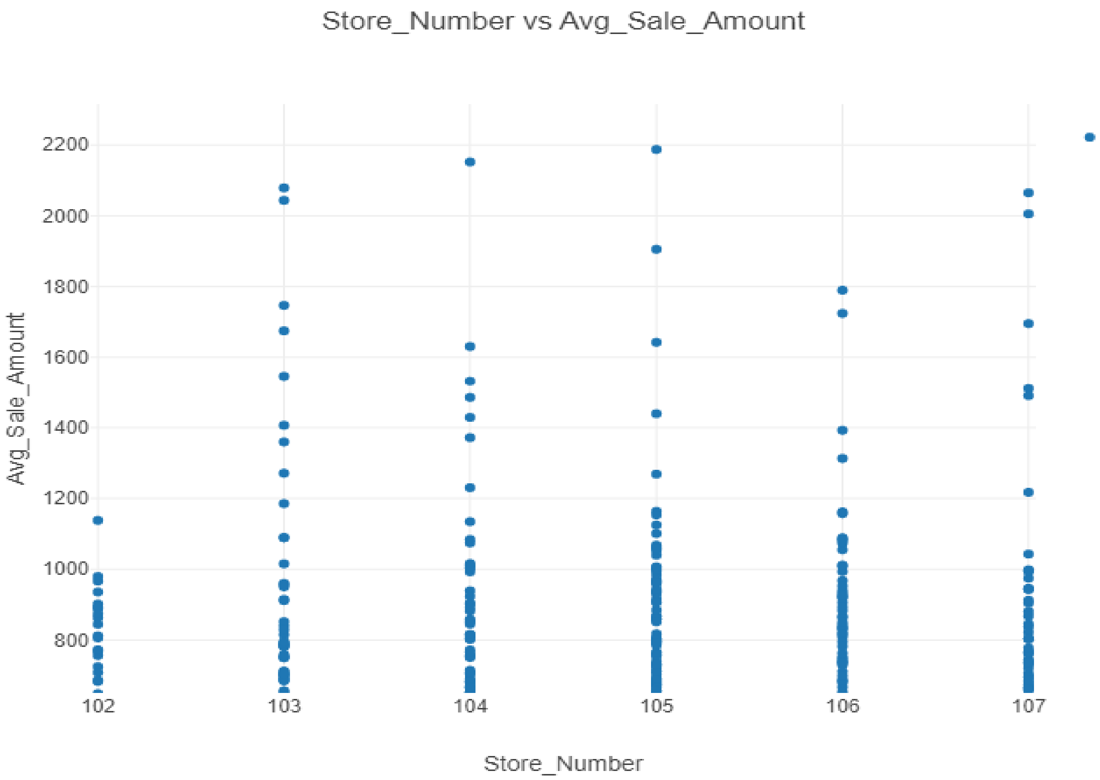


Fig. 3

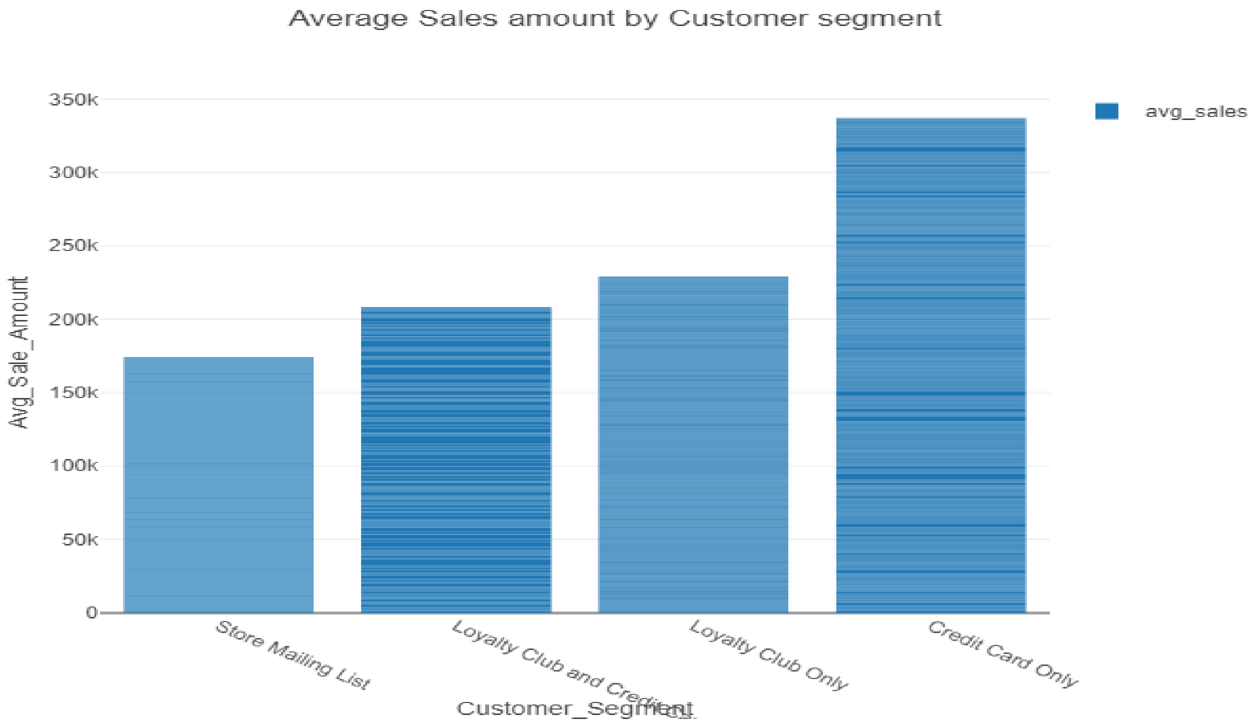
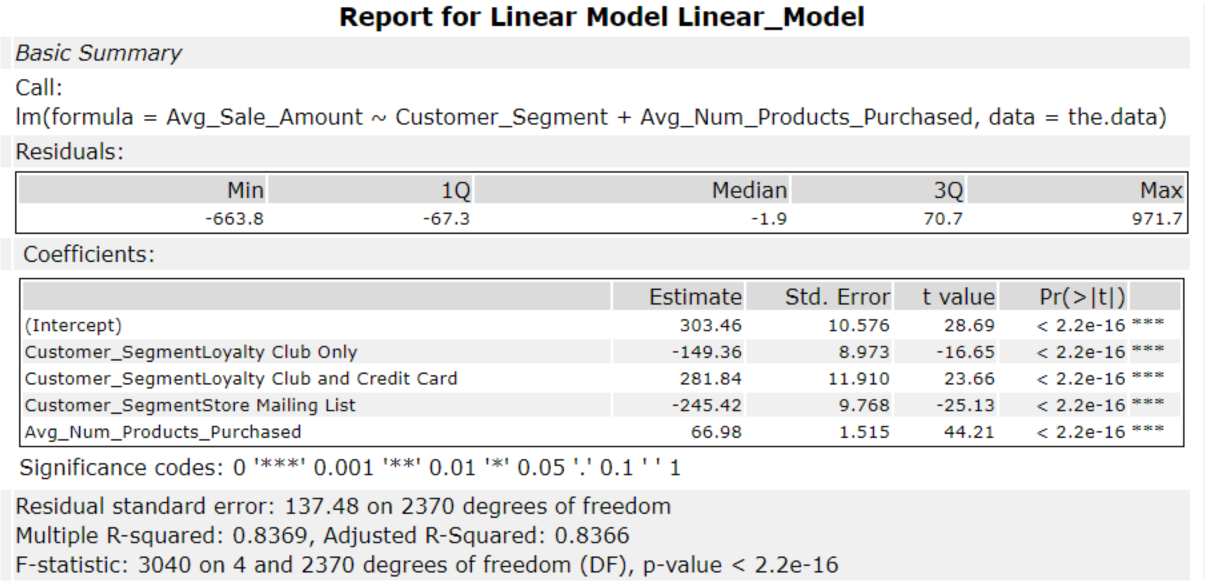


Fig. 4

The above graphs showed the correlations that existed between the predictor variables and the target variables. Since Customer_segment is a categorical variable I represented the contributions using a bar chart and fig.1 shows a very strong correlation with the target variable, that's is average number of sales.

2. Explain why you believe your linear model is a good model.

After running the model, below is the report that emerges. I will use this report and explain why my linear model is a good one.



Report for Linear Model Linear_Model

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

An R-Squared value for the Multiple linear regression was "0.8369" while the Adjusted R-Squared was "0.8366". The reduction shows the involvement of more variable that has less predictive power. The values from my model are good as it was confirmed by the P-value being far less than 0.05. The P-value shows that the model is more than 95% confident that there exists a relationship between the predictors and the target variable.

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)

Below is the nest linear regression equation based on the available data.

Avg_Sales_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) - 245.52 * (Customer_Segment: Store Mailing List) + 0 * (Customer_Segment: Credict Card Only)

Step 3: Presentation/Visualization

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

At the minimum, answer these questions:

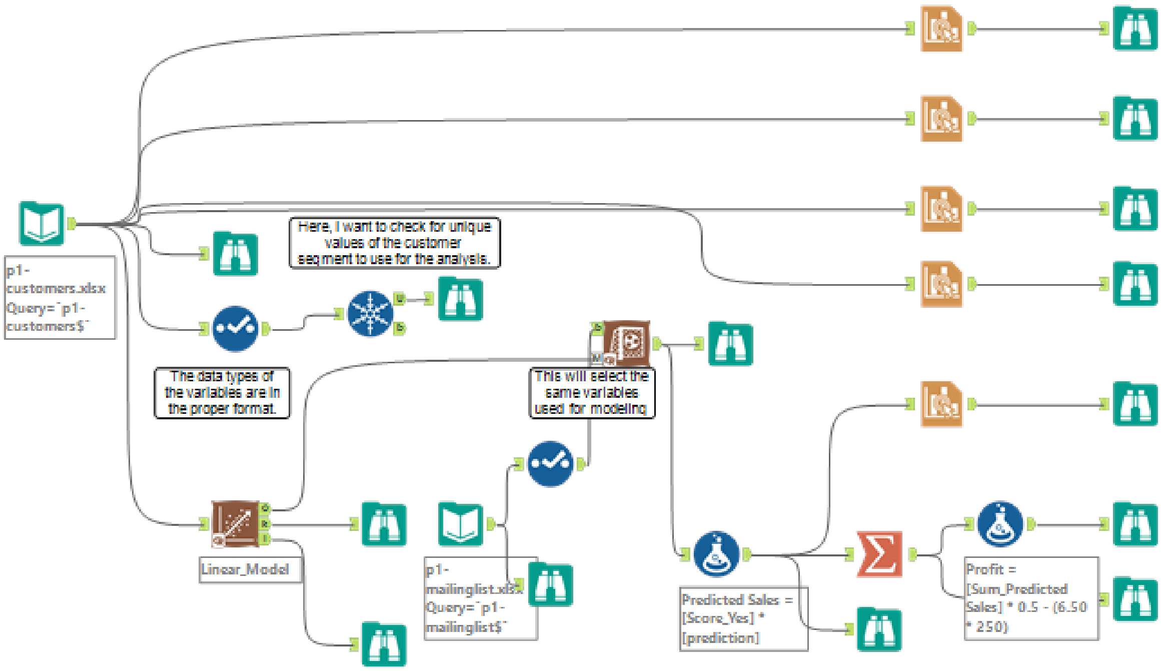
1. What is your recommendation? Should the company send the catalog to these 250 customers?

I recommend the company send the catalog to these 250 customers as more profit than expected will be made from them. This is the result obtained from the predictive analysis.

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

The company expected a profit contribution of $10,000 and my analysis shows they will make about $21,000 while considering the probability that these 250 customers will respond to the catalog that was sent to them. This profit contribution was obtained by taking 50% of the entire revenue and then subtracting $6.50 for each catalog sent to these customers.

Note: My Alteryx workflow is in the zip file in which this pdf file is included, but below is the view of the workflow.



Here, I want to check for unique values of the customer segment to use for the analysis.

The data types of the variables are in the proper format.

This will select the same variables used for modeling

p1-customers.xlsx Query=`p1-customers$`
Linear_Model
p1-mailinglist.xlsx Query=`p1-mailinglist$`
Predicted Sales = [Score_Yes] * [prediction]
Profit = [Sum_Predicted Sales] * 0.5 - (6.50 * 250)

Fig. 6 Alteryx Workflow

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

The expected profit is calculated below:

Profit = [Sum_Predicted Sales] * 0.5 - (cost of each catalog * 250)

= $21, 987.44

Assuming the catalog was sent to these customers, a profit contribution of $21, 987.44 is expected from them.