Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   The decision needed to be made is to determine whether company should send the new catalog to 250 new customers or not. The deciding factor will be whether these customers yield a minimum profit of $10,000 or not. The manager must decide based on the predictive model whether to send the new catalog or not.

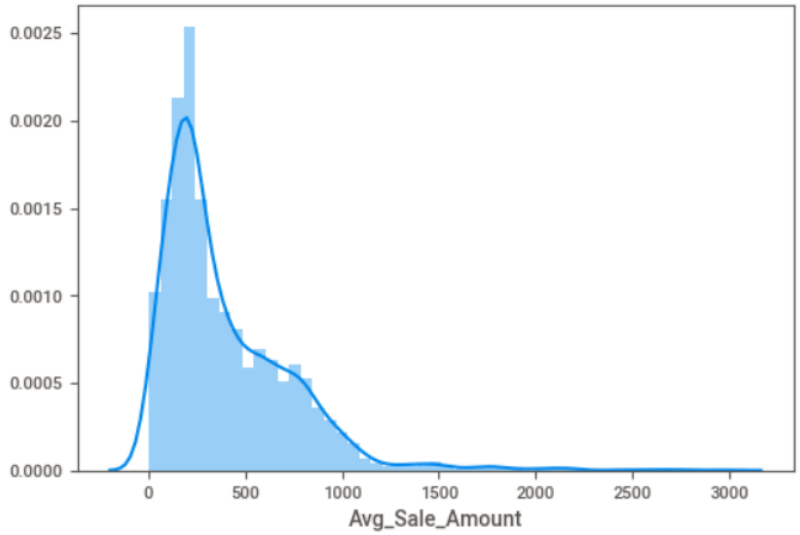
2. What data is needed to inform those decisions?
   The data needed to make these decisions is predictive average sales amount for all these new customers. This data will be obtained by applying linear regression and predicting average sales amount for these new customers. Linear regression will be applied on previous customers which will help us predict for these new customers and help us make decisions.

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)

To set up our linear regression model first we will have to know our Target variable which should be a numerical continuous variable.

Target Variable: Average Sales Amount.

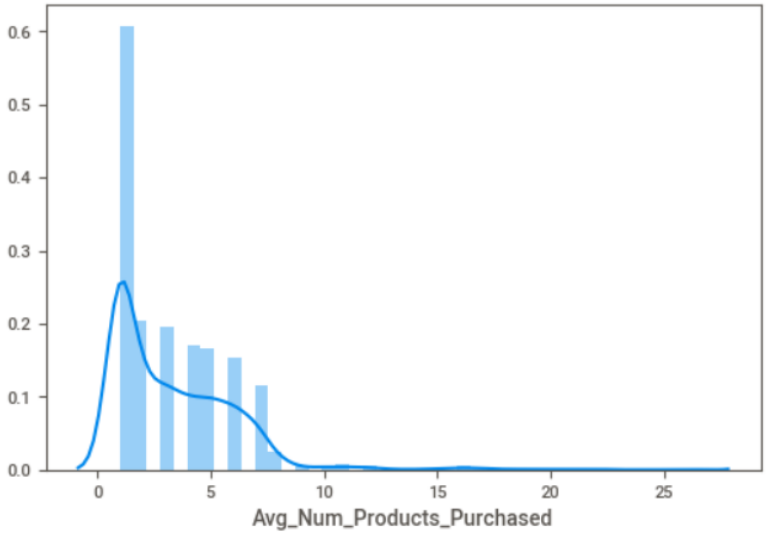


As we can clearly see that the distribution of the variable is skewed to the right. So, the data is not balanced we can go the manager and ask for more data or apply transformation methods which will help to bring the distribution approximately normal.

Now we will decide our predictor variables which will help to predict average sales amount for a customer. The variables such as Name, address and customer id will not be used for our analysis as they represent customer information which will not affect the sales amount. So, for predicting sales amount these variables will not be used.
Variable State has only one unique value which indicate that all the present data is from one state only. So, we will not include it in our analysis.

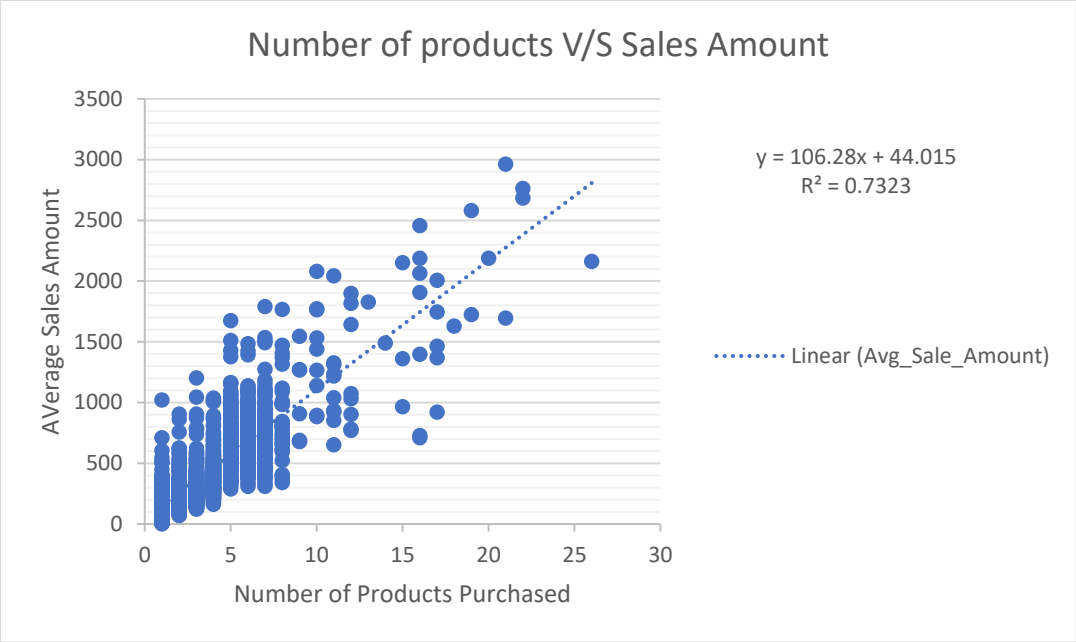
Continuous Variable available are:

1. **Average Number of Products Purchased.**
   Distribution of the variable:



   As we can clearly see that the distribution of the variable is skewed to the right. So, the data is not balanced we can go the manager and ask for more data or apply transformation methods.

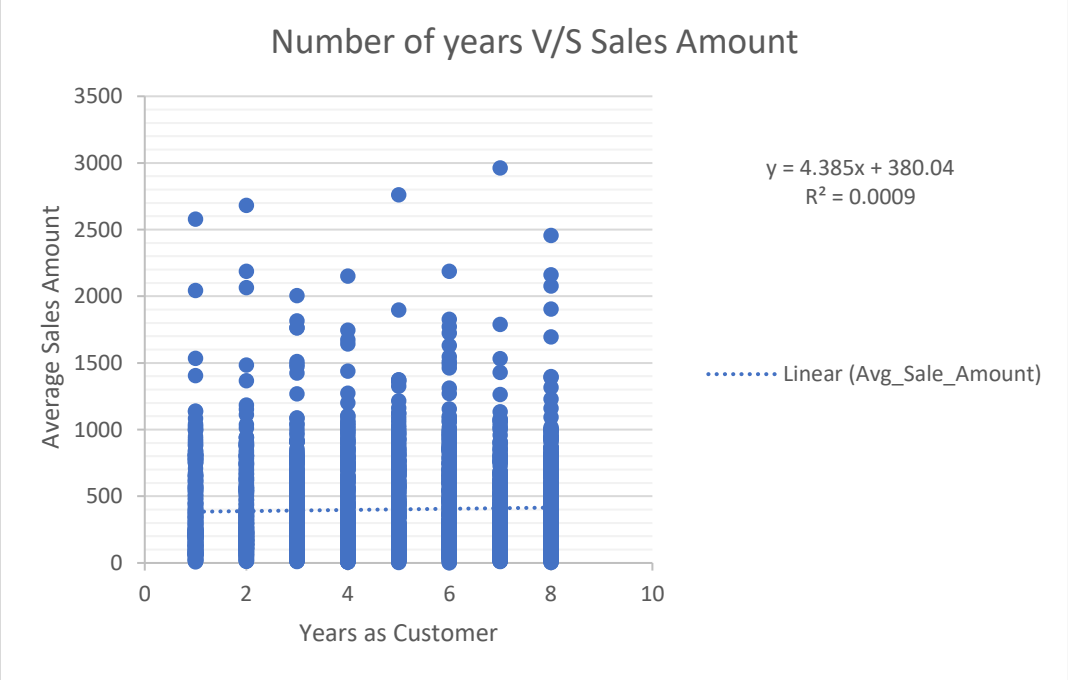
   Relationship with Target Variable:



The variable is having a strong positive relationship with the target variable so we will **include it in the predictors.**

2. **Number of Years as Customer**

Relationship with target variable



The variable is not having any kind of relationship with the target variable so we will **not include it in the predictors.**

The rest all variables are either categorical or nominal variable for them we will find out p-values if they are significant for our analysis or not.

Categorical Variables:

- **Customer Segment**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	682.7	8.354	81.72	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16	***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16	***

As the p-value is less than 0.05 for all the columns we will include this variable in the predictors.

- **Store Number**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	412.505	18.84	21.89137	< 2.2e-16	***
Store_Number101	-15.042	27.83	-0.54050	0.58891	
Store_Number102	-32.077	41.44	-0.77416	0.43891	
Store_Number103	-6.012	29.49	-0.20389	0.83846	
Store_Number104	-26.233	28.00	-0.93701	0.34885	
Store_Number105	6.735	27.10	0.24851	0.80376	
Store_Number106	-30.484	27.64	-1.10279	0.27023	
Store_Number107	1.497	29.45	0.05085	0.95945	
Store_Number108	-53.173	30.10	-1.76629	0.07748	.
Store_Number109	14.657	32.25	0.45450	0.64951	

As the p-value I more than 0.05 for all the columns we will not include this variable in the predictors as these are statistically insignificant predictors.

- **City and Zip Code**
  Both columns also have p-value more than 0.05 for all the columns so we will not include these variable as they are statistically insignificant.

Model Performance on Train data:

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

Here, we can clearly see that all the variables used are statistically significant as they have p-values less than 0.05.
The R^2 for our model is 83.69% which says our model predicts at least 83.69% of the data correctly which is a good estimation and we can trust this model.

The best linear regression equation based on the available data:

Y = 303.46 + 0 (if Credit Card only) – 149.36 (If Customer_SegmentLoyalty Club Only)

+ 281.84 (if Customer_SegmentLoyalty Club and Credit Card)
– 245.42 (if Customer_SegmentStore Mailing list)
+ (66.98 x 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 defiantly send the catalog to the new customers as the expected profit generated for the company would be around $21,987.44. So, as we can clearly see that it is more than $10,000 and we can trust our model as it predicts around 84% of the time correctly. So, we can show faith in our model and recommend company to send out catalog to these new customers.

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

The expected profit which our model predicted is $21,987,44 which is more than double the minimum expected profit requirement which was $10,000. The expected total profit was calculated using:
Total expected profit = sum (predicted profit * 0.50 – 6.50)
We multiply by 0.50 as the profit was 50% of the total revenue generated by each customer.

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

Expected Profit is $21987.44

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.