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 business decision here is to determine whether to send or not to send the Company catalogue to new customers on the business mailing list.

2. What data is needed to inform those decisions?

Answer: The data that is needed to make this decision is historical customer data containing sales information in previous cases where catalogues have been sent to customers.

Step 2: Analysis, Modeling, and Validation

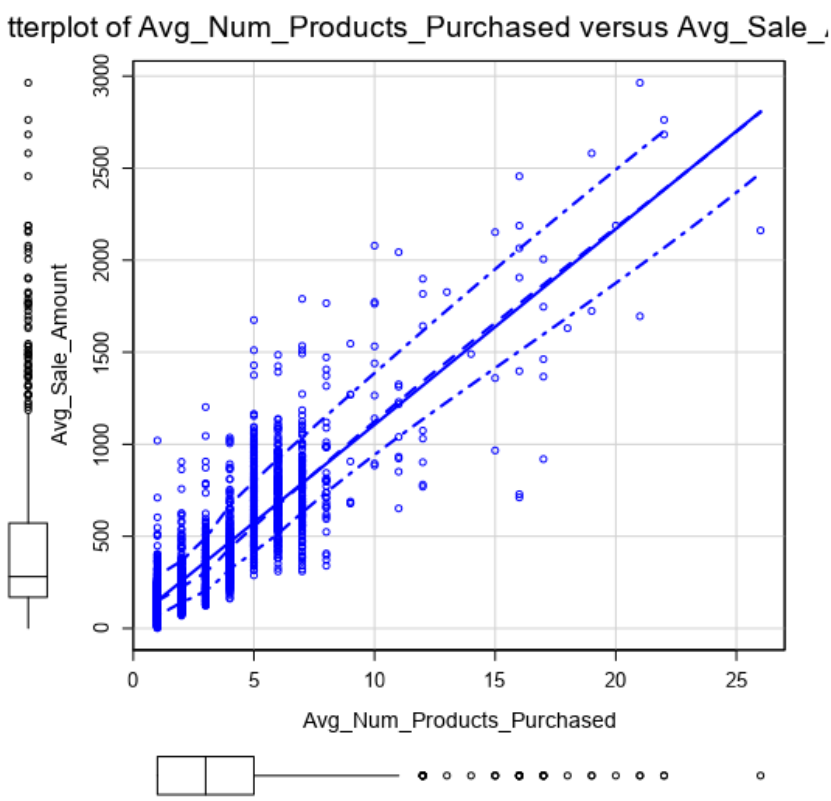
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

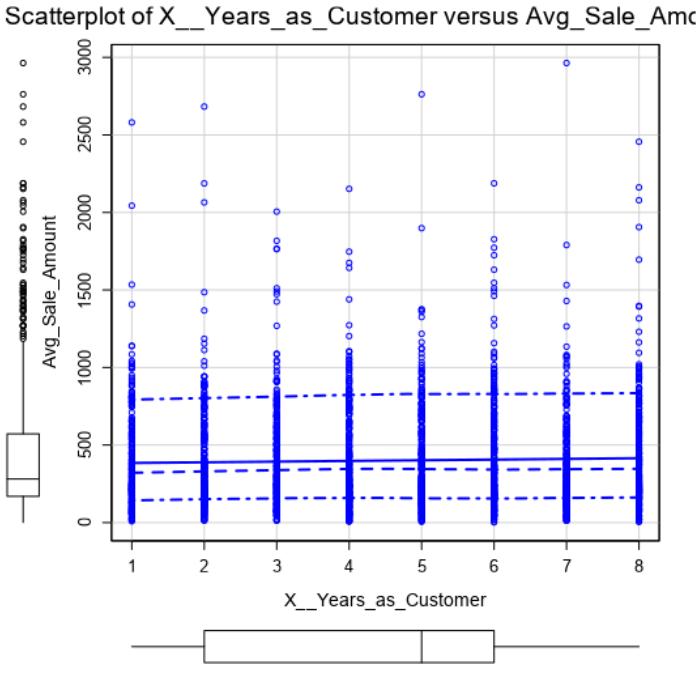
Important: Use the p1-customers.xlsx to train your linear model.

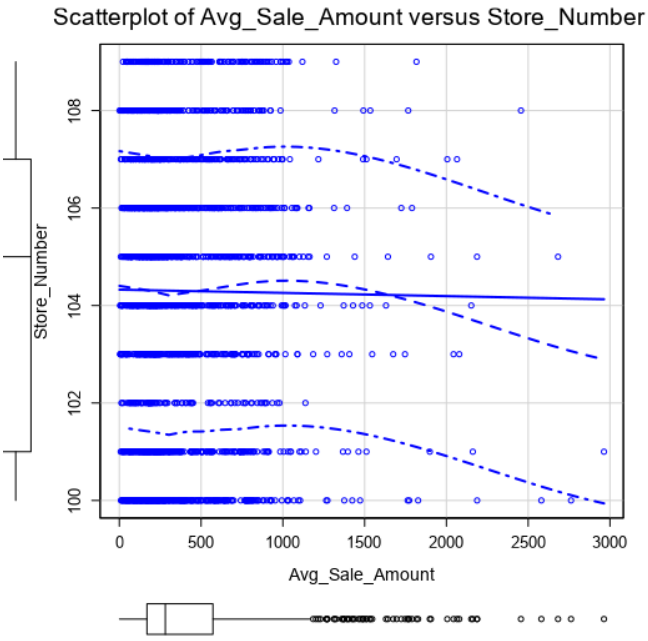
At the minimum, answer these questions:

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: I used two predictor variables for my model. The Average number of products and the Customer Segment; one categorical and the other numeric. The Average number of products was the only numerical feature that showed correlation and is linearly related with the target variable because the average sales amount increases as the average amount of products purchased increases as shown in below:


tterplot of Avg_Num_Products_Purchased versus Avg_Sale_
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased


Scatterplot of X__Years_as_Customer versus Avg_Sale_Amc
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
X__Years_as_Customer



The other feature in the model is the Categorical variable: Customer segment, I picked this feature for my model after several iterations of the model because it is the most statistically significant among all the categorical variables as shown below:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	493.3116	122.893	4.01416	6e-05	***
Customer_SegmentLoyalty Club Only	-150.6401	9.014	-16.71251	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.7801	11.956	23.65203	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.8152	9.888	-24.55721	< 2.2e-16	***
CityAurora	-20.5819	11.086	-1.85660	0.06349	.
CityBoulder	-41.1805	80.029	-0.51457	0.6069	
CityBrighton	-59.4890	97.639	-0.60927	0.5424	
CityBroomfield	-4.3414	15.124	-0.28705	0.7741	
CityCastle Pines	-93.0347	97.642	-0.95282	0.34078	
CityCentennial	-9.5731	18.158	-0.52721	0.59809	
CityCommerce City	-33.2255	44.454	-0.74742	0.45489	
CityDenver	0.2317	10.551	0.02196	0.98248	
CityEdgewater	27.9712	40.612	0.68875	0.49105	
CityEnglewood	6.0143	20.737	0.29002	0.77183	
CityGolden	-11.4221	32.719	-0.34910	0.72705	
CityGreenwood Village	-44.4576	38.059	-1.16812	0.24288	
CityHenderson	-285.8339	137.847	-2.07357	0.03823	*
CityHighlands Ranch	-28.1976	30.420	-0.92694	0.35405	
CityLafayette	-43.7104	62.140	-0.70342	0.48186	
CityLakewood	-7.3541	12.858	-0.57195	0.56741	
CityLittleton	-28.7184	18.967	-1.51412	0.13013	
CityLone Tree	77.3956	137.769	0.56178	0.57432	
CityLouisville	-30.5955	69.266	-0.44171	0.65874	
CityMorrison	-18.6190	52.789	-0.35271	0.72434	
CityNorthglenn	-14.7157	29.393	-0.50066	0.61666	

CityEdgewater	27.9712	40.612	0.68875	0.49105	
CityEnglewood	6.0143	20.737	0.29002	0.77183	
CityGolden	-11.4221	32.719	-0.34910	0.72705	
CityGreenwood Village	-44.4576	38.059	-1.16812	0.24288	
CityHenderson	-285.8339	137.847	-2.07357	0.03823	*
CityHighlands Ranch	-28.1976	30.420	-0.92694	0.35405	
CityLafayette	-43.7104	62.140	-0.70342	0.48186	
CityLakewood	-7.3541	12.858	-0.57195	0.56741	
CityLittleton	-28.7184	18.967	-1.51412	0.13013	
CityLone Tree	77.3956	137.769	0.56178	0.57432	
CityLouisville	-30.5955	69.266	-0.44171	0.65874	
CityMorrison	-18.6190	52.789	-0.35271	0.72434	
CityNorthglenn	-14.7157	29.393	-0.50066	0.61666	
CityParker	-6.0965	28.177	-0.21636	0.82873	
CitySuperior	-56.1322	46.681	-1.20245	0.22931	
CityThornton	29.0992	24.814	1.17271	0.24103	
CityWestminster	-6.6966	17.284	-0.38745	0.69846	
CityWheat Ridge	8.9128	20.673	0.43114	0.66641	
Store_Number	-1.6365	1.146	-1.42779	0.15348	
Responded_to_Last_CatalogYes	-29.5786	11.335	-2.60943	0.00913	**
Avg_Num_Products_Purchased	66.9147	1.527	43.81327	< 2.2e-16	***
Years_as_a_Customer	-2.3411	1.231	-1.90197	0.0573	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.22 on 2341 degrees of freedom

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: I believe my model is a good model because the R-squared value at 0.8369 is high and the predictor variable has p-value less than 0.5 which implies that it is statistically significant. A high R squared and a p-value shows that the model is a good fit for the data.

	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

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)

Answer: Y = 303.46 – 149.36*(If Customer_SegmentLoyalty: Club Only) + 281*(If Customer_SegmentLoyalty: Club and Credit Card) – 245.42*(If Customer_Segment: Mailing List) + 66.98*(Average number of products purchased)

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?

Recommendation: I will recommend that the company send the catalogue to the 250 customers as the results from my model shows that it will be profitable for the company to do so.

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

Answer: Using data from the previous time the catalogue was sent, I trained a linear regression model and use it to predict the sales amount if the catalogue was sent to the people on the mailing list. And I calculated the gross margin and removed the cost of the catalogues from the results.

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

Answer: The expected profit from the new catalogue from my calculations is $21,987.