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?

Decision is to whether or not to print and ship catalogs to 250 new customers.

Criteria is whether the predicted profits from the 250 new customers exceed $10,000. This is after considering the cost of printing and distributing each catalog at $6.50 as well as the average 50% gross margin from each product sold.

2. What data is needed to inform those decisions?

Need to predict how much sales can be made from each new customer based on historical customer data such as previous sales amount, number of products sold, segment or type of customer (loyalty card, credit card, mailing list), number of years as customer and geospatial location.

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.

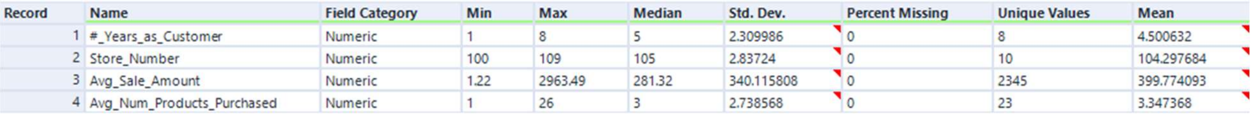
At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model?

Data exploration is first performed to understand the distribution of the variables and their relationship with the target, ‘Avg_Sale_Amount’. Unique variables, ‘Name’, ‘Customer_ID’ and ‘Address’ are dropped, as is ‘City’.

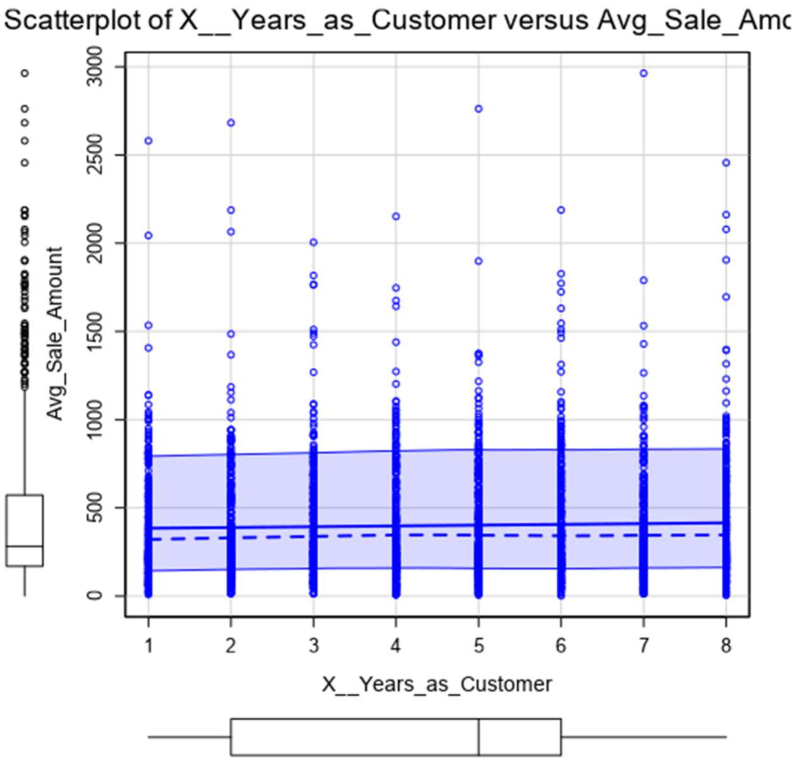
An initial linear regression model was fitted using the remaining variables, ‘Customer_Segment’, ‘ZIP’, ‘Store Number’, ‘Avg_Num_Products_Purchased’, ‘#_Years_as_Customer’ against the target of ‘Avg_Sale’Amount’ to calculate the strength of each of their relationship through their p values.

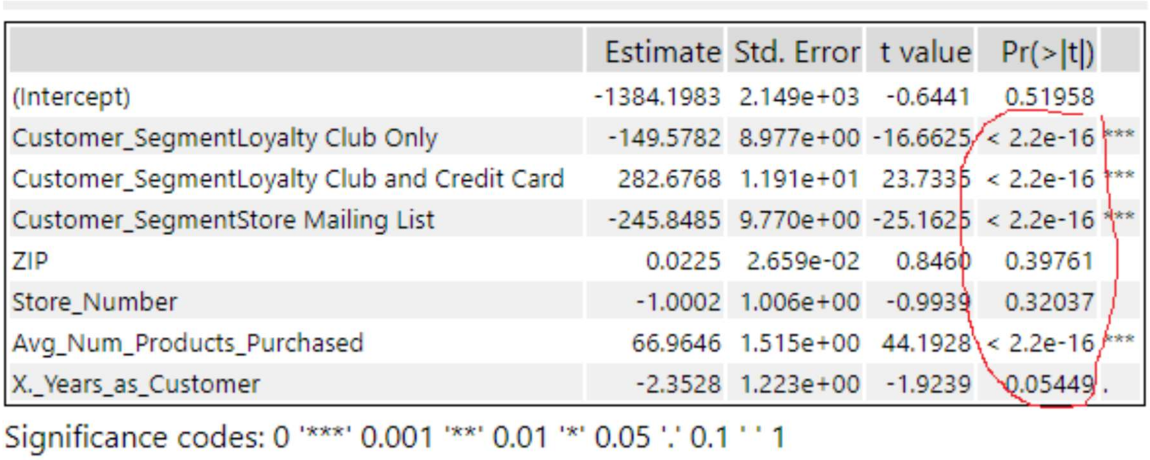
Numeric variables includes ‘#_Years_as_Customer’, ‘Avg_Sale_Amount’, ‘Store_Number’ and ‘Avg_Num_Products_Purchased’. The overall statistics are shown below.



Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	#_Years_as_Customer	Numeric	1	8	5	2.309986	0	8	4.500632
2	Store_Number	Numeric	100	109	105	2.83724	0	10	104.297684
3	Avg_Sale_Amount	Numeric	1.22	2963.49	281.32	340.115808	0	2345	399.774093
4	Avg_Num_Products_Purchased	Numeric	1	26	3	2.738568	0	23	3.347368

The scatterplot between ‘#_Years_as_Customer’ and ‘Avg_Sale_Amount’ shows a rather poor relationship. This is supported by a p value of 0.05825 which indicates a weak correlation according to significance codes.

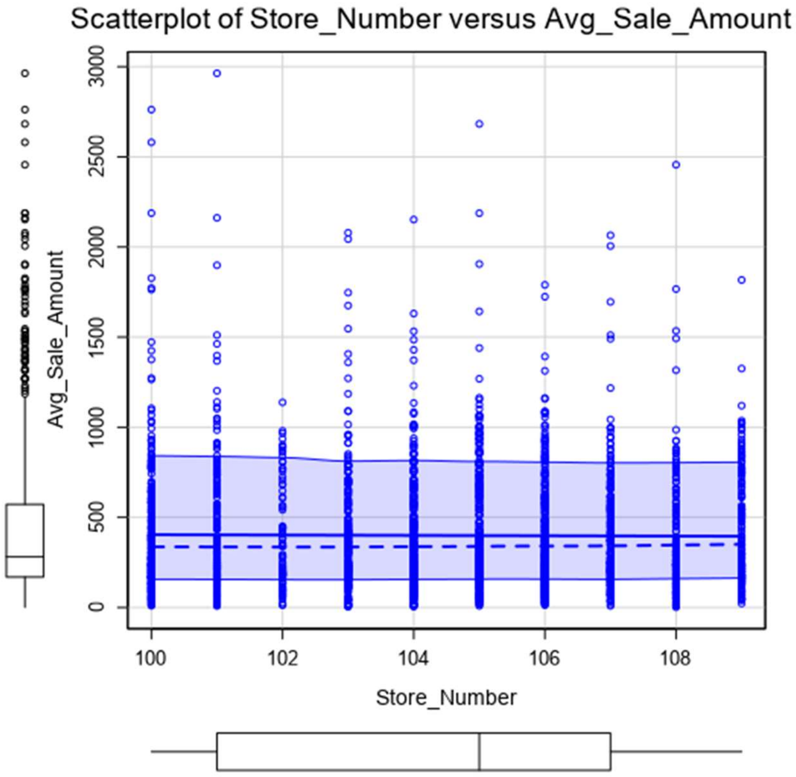




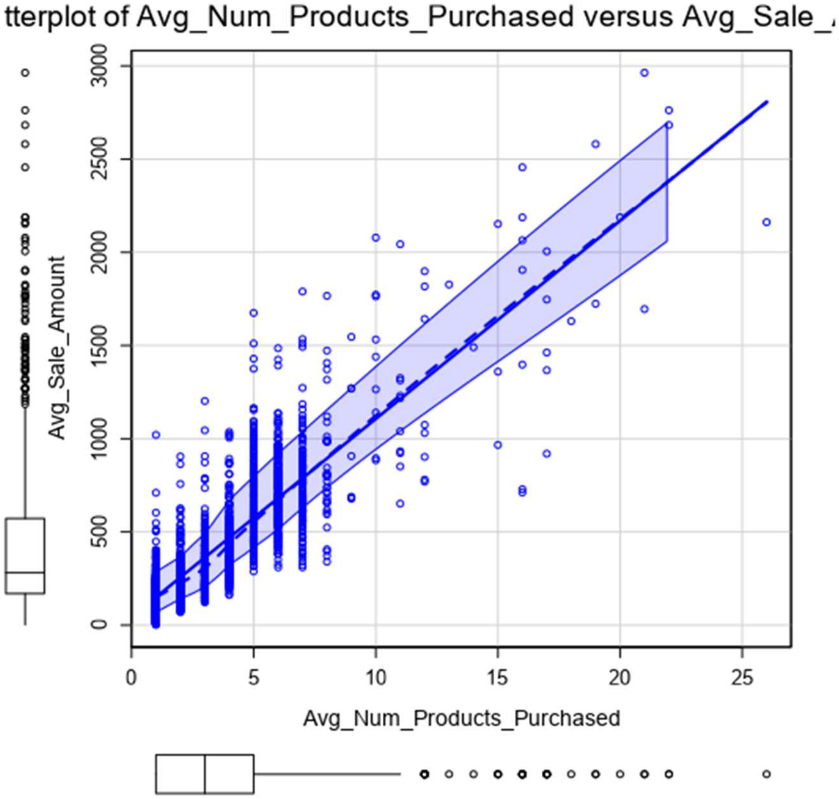
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958	
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00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

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

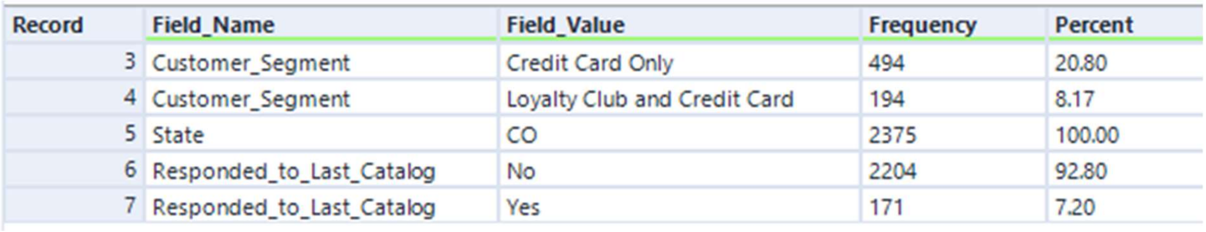
Similarly, for ‘Store_Number’, the scatterplot with ‘Avg_Sale_Amount’ also shows a rather poor relationship and is supported by a p value of 0.32036 which indicates a weak correlation as it is greater 0.05.



For ‘Avg_Num_Products_Purchased’, the relationship with ‘Avg_Sale_Amount’ is strongly linear based on the scatterplot and a very low p value of <2.2e-16.



For non-numeric or categorical variables, ‘Customer_Segment’, ‘State’ and ’Responded_to_Last_catalog’, the overall statistics are shown below. Since all previous data are from the same state, CO, this variable is not useful and is dropped. ‘Responded_to_Last_Catalog’ is also dropped since it is not available in the dataset for the new customers.

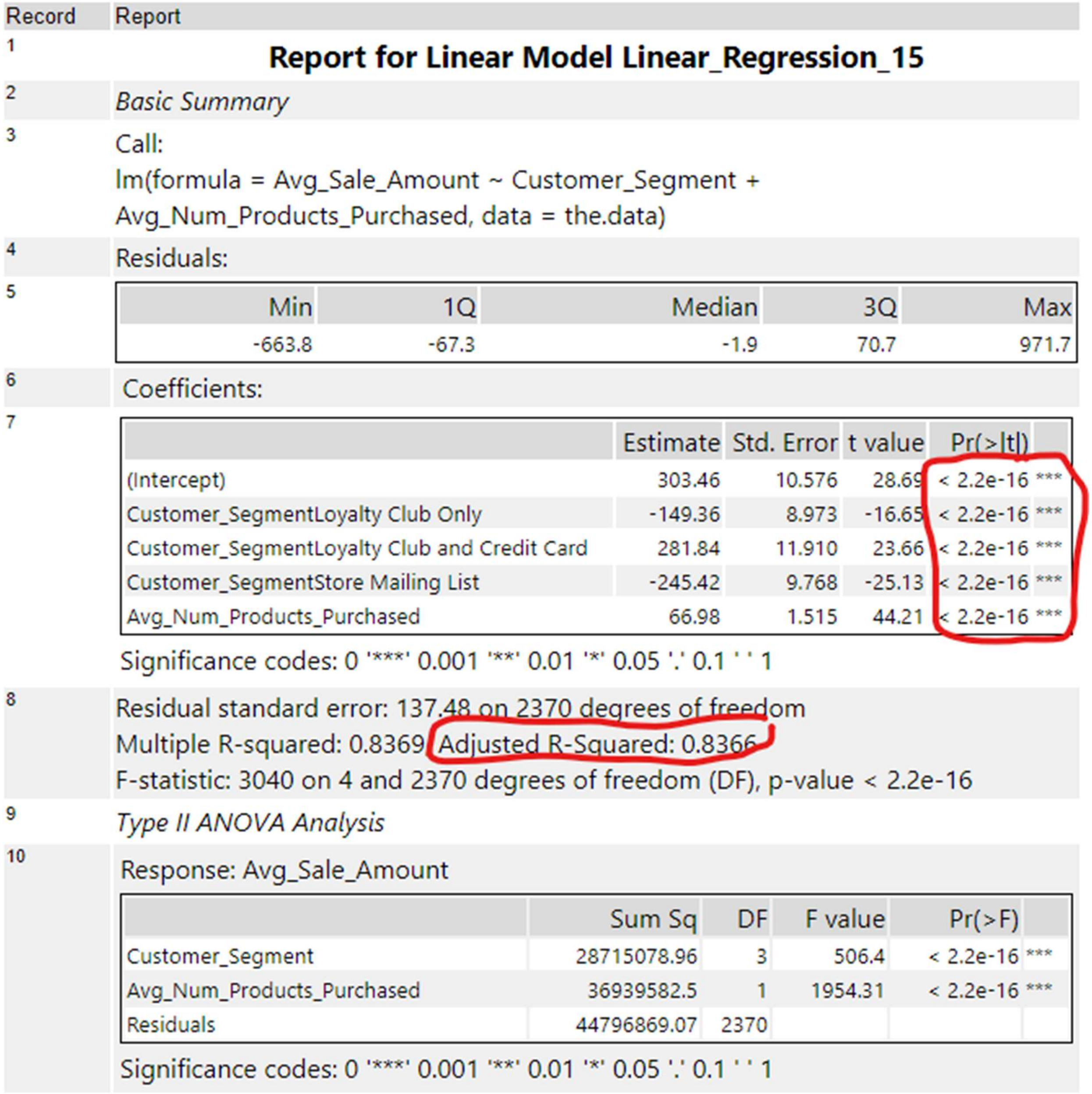


Record	Field_Name	Field_Value	Frequency	Percent
3	Customer_Segment	Credit Card Only	494	20.80
4	Customer_Segment	Loyalty Club and Credit Card	194	8.17
5	State	CO	2375	100.00
6	Responded_to_Last_Catalog	No	2204	92.80
7	Responded_to_Last_Catalog	Yes	171	7.20

The final variables, which have strong correlations with ‘Avg_Sale_Amount’, selected are ‘Customer_Segment’, ‘Avg_Num_Products_Purchased’.

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

Fitting ‘Customer_Segment’ and ‘Avg_Num_Products_Purchased’ to ‘Avg_Sale_Amount’ with a multilinear regression model resulted in a strong correlated model with a high adjusted R2 of 0.8366 and very small p values of < 2.2e-16 for each coefficient as shown in summary below.



Record	Report
1	**Report for Linear Model Linear_Regression_15**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	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

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)

Avg_Sale_Amount = 303.46
– 149.36 * Customer_Segment(Loyalty Club)
+ 281.84 * Customer_Segment(Loyalty Club and Credit Card)
– 245.42 * Customer_Segment(Mailing List)
+ 66.98 * Avg_Num_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?

It is recommended to print and ship the catalogs to the 250 new customers as the expected profit is predicted to be greater than $10,000.

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

The linear regression model consists of 2 variables, one categorical (Customer_Segment) and the other is numerical (Avg_Num_Producted_Purchased).

For each new customer, the linear regression model was applied to predict their sales amount, which was then multiplied with the probability of sales (‘Score_Yes’) to obtain the expected sales.

The 50% gross margin was then applied to the probable sales to obtain the gross. The cost due to printing and shipment for each catalog was offset by subtracting the $6.50 to get the profit for each new customer. Summing up the net profit for all 250 customer gives the total net profit which is greater than the $10,000 threshold.

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

Expected profit for 250 new customers is $21,987.44.