Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

This report predicts how much money the company can expect to earn from sending out catalogs to new customers by building the model and applying the results in order to provide a recommendation to the management.

Key Decisions:

1. **Decisions which need to be made**

   We need to determine how much profit the company can expect from sending a catalog to each of 250 new customers from the company's mailing list, and predict whether the expected profit from these 250 new customers exceeds $10,000. If the expected profit exceeds $10,000, the company will send the catalog out to each of these new customers.

2. **Data needed to inform those decisions**

   Considering the data we had, I hypothesised that the following variables might be useful for making the prediction:
   - Customer_Segment
   - City
   - ZIP
   - Store_Number
   - Avg_Num_Products_Purchased
   - #_Years_as_Customer (I changed the name to Num_Years_as_Customer)

   Our predictor variables should be variables which can be found in both the dataset on which we build our model and the dataset from which we need to predict sales. They should also contain more than one unique record.

   Our target variable is Avg_Sale_Amount.

Step 2: Analysis, Modeling, and Validation

1. How and why I selected the predictor variables

I used scatter plots between an individual variable and the target variable for numerical variables to see if a variable might be a good candidate for a predictor variable. I checked the p-values of categorical and numerical variables to see which variables were statistically significant (p-value <= 0.05).

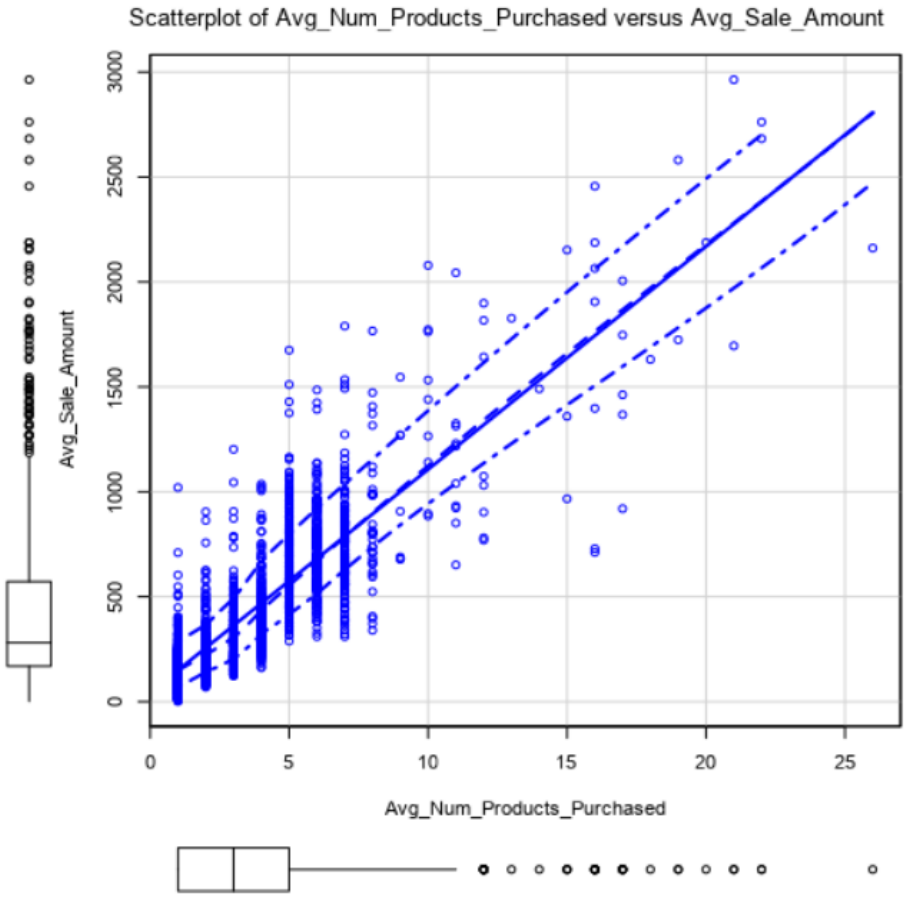


Figure 2.1.1: Avg_Num_Products_Purchased versus Avg_Sale_Amount

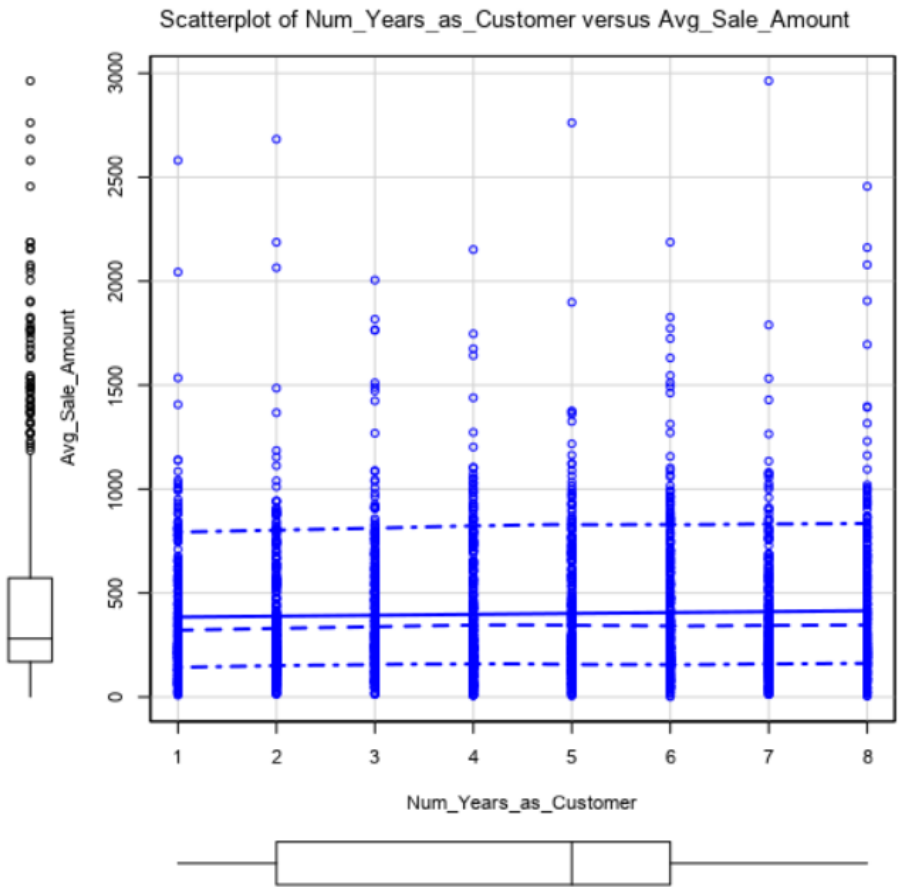


Figure 2.1.2: Num_Years_as_Customer versus Avg_Sale_Amount

Figure 2.1.1 shows a positive slope, indicating as the average number of products purchased increases the average sale amount also increases, and the two variables are positively related. On the other hand, figure 2.1.2 shows a minuscule, positive slope and the data points are quite spread out, indicating there might well be an extremely weak relationship between the number of years since the customer's first purchase and the average sale amount. This might indicate the average number of products purchased is a good predictor variable for the average sale amount, while customer tenure is not.

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

Min	1Q	Median	3Q	Max
-1033.2	-99.4	-17.8	71.4	1099.4

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	44.02	5.704	7.716	1.75e-14	***
Avg_Num_Products_Purchased	106.28	1.319	80.572	< 2.2e-16	***

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

Residual standard error: 176.01 on 2373 degrees of freedom
Multiple R-squared: 0.7323, Adjusted R-Squared: 0.7322
F-statistic: 6492 on 1 and 2373 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Avg_Num_Products_Purchased	201109435.07	1	6491.91	< 2.2e-16	***
Residuals	73511948.03	2373			

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

Figure 2.1.3: Avg_Num_Products_Purchased versus Avg_Sale_Amount

Record Report

1

Report for Linear Model
Linear_Regression_Catalog_Demand__Num_Years_as_Customer

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ Num_Years_as_Customer, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-410	-233	-119	174	2553

6 Coefficients:

7

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	380.039	15.283	24.867	< 2.2e-16	***
Num_Years_as_Customer	4.385	3.021	1.451	0.14679	

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

8 Residual standard error: 340.04 on 2373 degrees of freedom
Multiple R-squared: 0.000887, Adjusted R-Squared: 0.0004659
F-statistic: 2.107 on 1 and 2373 degrees of freedom (DF), p-value 0.1468

9 *Type II ANOVA Analysis*

10 Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Num_Years_as_Customer	243578.02	1	2.11	0.14679
Residuals	274377805.08	2373		

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

Figure 2.1.4: Num_Years_as_Customer versus Avg_Sale_Amount

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

Min	1Q	Median	3Q	Max
-1001.85	-71.66	3.08	73.02	1889.33

Coefficients:

	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	***

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

Residual standard error: 185.67 on 2371 degrees of freedom
Multiple R-squared: 0.7024, Adjusted R-Squared: 0.702
F-statistic: 1865 on 3 and 2371 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	192884931.52	3	1865.06	< 2.2e-16	***
Residuals	81736451.57	2371			

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

Figure 2.1.5: Customer_Segment versus Avg_Sale_Amount

Record	Report
1	**Report for Linear Model** **Linear_Regression_Catalog_Demand__City**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ City, data = the.data)
4	Residuals:

Min	1Q	Median	3Q	Max
-560	-232	-116	175	2559

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	386.087	21.67	17.81399	< 2.2e-16	***
CityAurora	18.755	26.55	0.70630	0.48007	
CityBoulder	154.103	197.85	0.77890	0.43612	
CityBrighton	-291.157	241.83	-1.20398	0.22872	
CityBroomfield	7.409	37.39	0.19816	0.84294	
CityCastle Pines	-193.877	241.83	-0.80171	0.4228	
CityCentennial	-13.816	44.24	-0.31230	0.75484	
CityCommerce City	296.728	109.87	2.70065	0.00697	**
CityDenver	18.551	24.99	0.74237	0.45794	
CityEdgewater	76.875	100.69	0.76349	0.44525	
CityEnglewood	-9.806	50.41	-0.19450	0.8458	
CityGolden	-12.719	81.09	-0.15685	0.87538	
CityGreenwood Village	-60.038	93.58	-0.64157	0.52121	
CityHenderson	-171.697	341.31	-0.50305	0.61498	
CityHighlands Ranch	4.904	74.26	0.06604	0.94735	
CityLafayette	-41.955	153.86	-0.27267	0.78513	
CityLakewood	31.652	31.69	0.99872	0.31803	
CityLittleton	-9.727	45.62	-0.21322	0.83118	
CityLone Tree	468.783	341.31	1.37348	0.16973	
CityLouisville	-37.619	171.68	-0.21912	0.82658	
CityMorrison	126.608	130.55	0.96977	0.33226	
CityNorthglenn	-29.332	72.83	-0.40276	0.68716	
CityParker	-51.059	69.04	-0.73953	0.45966	
CitySuperior	-81.067	115.59	-0.70133	0.48317	
CityThornton	8.199	61.52	0.13327	0.89399	
CityWestminster	9.430	42.83	0.22016	0.82576	
CityWheat Ridge	43.875	51.17	0.85744	0.39129	

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

8 Residual standard error: 340.62 on 2348 degrees of freedom
Multiple R-squared: 0.008008, Adjusted R-Squared: -0.002976
F-statistic: 0.7291 on 26 and 2348 degrees of freedom (DF), p-value 0.8374

9 *Type II ANOVA Analysis*

10 Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
City	2199299.15	26	0.73	0.83744
Residuals	272422083.94	2348		

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

Figure 2.1.6: City versus Avg_Sale_Amount

Record	Report
1	**Report for Linear Model Linear_Regression_Catalog_Demand__ZIP**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ ZIP, data = the.data)
4	Residuals:

Min	1Q	Median	3Q	Max
-560	-229	-101	170	2516

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	381.7723	61.22	6.23582	5.33e-10	***
ZIP80003	29.6849	76.31	0.38901	0.69731	
ZIP80004	-16.6265	72.12	-0.23055	0.81768	
ZIP80005	13.7187	73.38	0.18695	0.85171	
ZIP80007	-14.5165	142.64	-0.10177	0.91895	
ZIP80010	74.0606	88.87	0.83335	0.40473	
ZIP80011	-7.8434	80.71	-0.09718	0.9226	
ZIP80012	51.6589	71.19	0.72561	0.46815	
ZIP80013	6.3222	68.29	0.09258	0.92624	
ZIP80014	67.5356	73.38	0.92035	0.35749	
ZIP80015	-0.7864	70.99	-0.01108	0.99116	
ZIP80016	1.6900	81.57	0.02072	0.98347	
ZIP80017	-47.3341	77.07	-0.61413	0.53919	
ZIP80018	149.4544	206.11	0.72513	0.46844	
ZIP80020	8.3685	71.30	0.11737	0.90658	
ZIP80021	28.3244	78.54	0.36063	0.71841	
ZIP80022	301.0427	123.97	2.42843	0.01524	*
ZIP80023	31.5877	123.97	0.25481	0.79889	
ZIP80026	-37.6403	164.28	-0.22913	0.81879	
ZIP80027	-63.3838	112.63	-0.56275	0.57366	
ZIP80030	-58.1050	119.63	-0.48571	0.62722	
ZIP80031	12.1337	77.35	0.15687	0.87536	
ZIP80033	28.6921	79.93	0.35896	0.71966	
ZIP80108	-189.5623	248.69	-0.76225	0.44599	
ZIP80110	-14.5138	99.32	-0.14614	0.88383	
ZIP80111	-28.7132	96.34	-0.29804	0.7657	
ZIP80112	-21.0100	101.01	-0.20800	0.83525	
ZIP80113	20.6973	93.81	0.22063	0.8254	
ZIP80120	-27.3769	107.21	-0.25535	0.79847	
ZIP80121	102.4527	123.97	0.82646	0.40863	
ZIP80122	-37.6673	123.97	-0.30385	0.76127	

S

8 Residual standard error: 340.87 on 2289 degrees of freedom
Multiple R-squared: 0.03151, Adjusted R-Squared: -0.004454
F-statistic: 0.8761 on 85 and 2289 degrees of freedom (DF), p-value 0.7824

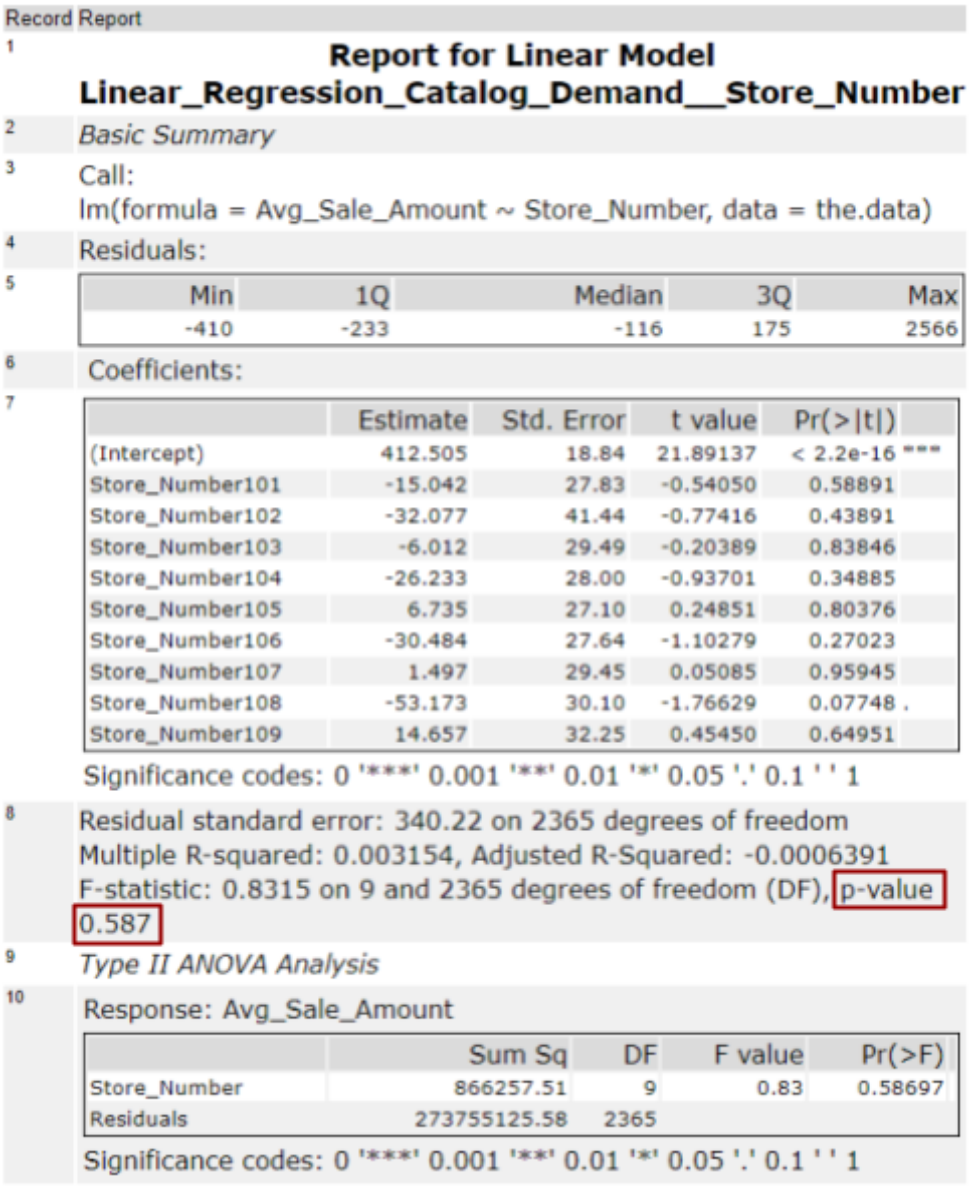
9 *Type II ANOVA Analysis*

10 Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
ZIP	8653238.79	85	0.88	0.78242
Residuals	265968144.31	2289		

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

Figure 2.1.7: ZIP versus Avg_Sale_Amount



Record Report

Report for Linear Model
Linear_Regression_Catalog_Demand__Store_Number

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Store_Number, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-410	-233	-116	175	2566

Coefficients:

	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	

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

Residual standard error: 340.22 on 2365 degrees of freedom
Multiple R-squared: 0.003154, Adjusted R-Squared: -0.0006391
F-statistic: 0.8315 on 9 and 2365 degrees of freedom (DF), p-value 0.587

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Store_Number	866257.51	9	0.83	0.58697
Residuals	273755125.58	2365		

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

Figure 2.1.8: Store_Number versus Avg_Sale_Amount

Figures 2.1.3 and 2.1.5 show Avg_Num_Products_Purchased and Customer_Segment are statistically significant as they both have p-values of < 2.2e-16 (i.e. there is a < 2.2e-14% chance the observed difference could have occurred by chance). In contrast, Figures 2.1.4, 2.1.6, 2.1.7 and 2.1.8 show Num_Years_as_Customer, City, ZIP and Store_Number are not statistically significant as they have p-values of 0.14679, 0.8374, 0.7824 and 0.587 respectively (i.e. there are 14.679%, 83.74%, 78.24% and 58.7% chances respectively the observed difference could have occurred by chance). Therefore, the average number of products purchased and customer segment are good predictor variables for the average sale amount, whereas the customer tenure, city, ZIP and store number are not.

2. Why I believe our linear model is a good model

Record	Report
1	**Report for Linear Model** **Linear_Regression_Catalog_Demand__Final**
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

Figure 2.2.1: Customer_Segment and Avg_Num_Products_Purchased versus Avg_Sale_Amount

Figure 2.2.1 shows we have improved the model with Customer_Segment and Avg_Num_Products_Purchased as Multiple R-squared and Adjusted R-Squared values are larger compared to those in Figures 2.1.3 and 2.1.5. It is better to rely on the adjusted R-squared value as it increases only if the new predictor added improves the model more than would be expected by chance. 83.66% of the variance for the target variable is explained by the predictor variables.

3. The best linear regression equation based on the available data

Y = 303.46 + 66.98 * Avg_Num_Products_Purchased - 149.36 (if Customer_Segment: Loyalty Club Only) + 281.84 (if Customer_Segment: Loyalty Club and Credit Card) - 245.42 (if Customer_Segment: Store Mailing List) + 0 (if Customer_Segment: Credit Card Only)

Step 3: Presentation/Visualization

1. **Recommendation (whether the company should send the catalogs to the 250 customers)**

   It is recommended that the company send the catalog to the new 250 customers as the expected profit exceeds $10,000.

2. **How I came up with the recommendation**

   I built the model on the dataset which contained information on our 2,375 existing customers (p1-customers.xlsx) and applied this model to the dataset which contained our 250 new customers (p1-mailinglist.xlsx) to obtain the predicted sale amount. After that, to obtain the expected sale amount, I multiplied the predicted sale amount by Score_Yes. For example, if our customer A Giametti is to buy from us, we predict this customer will buy $355.04 worth of products. At a 30.50% chance that this customer will actually buy from us, we can expect revenue to be $355.04 * 30.50% = $108.30 (all numbers are rounded to 2 decimal places). Next, I multiplied the expected sale amount by the average gross margin (50%) and then subtracted the cost of printing and distributing ($6.50) from it to obtain the expected profit per customer. (The gross margin calculation has not taken into account the cost of printing and distributing.) Finally, I added up all of the expected profit from each new customer to obtain the total expected profit.

3. **The expected profit from the new catalog (assuming the catalog is sent to the 250 customers)**

   The expected profit from the new catalog is $21,987.44. This number is twice the minimum expected profit of $10,000.