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.)

We have checked the data type of variable and see the what is type of data i,e continues and discrete. Analyze the what could be the predictor variable and target variable. By looking into data, it is been observed that Avg_sale_amount varies as per zip code, store_num , customer segment ,Avg_num_of_product_purchased, #_year_as_customer. Every zip code contains multiple stores having multiple customers who has avg_num_of product and #yr_as_customer. Sale_amount varies as per city but city is identified with the help of zip code.Customer segment contains various type such as **Credit Card Only, Loyalty Club Only, Loyalty Club and Credit Card, Store Mailing List**

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**
   - Catalogs will only be sent if the sum of expected profit exceeds $10,000.

   As we can see that sum of expected profit 21987.43 which exceeds $10,000, So **Company should send catalog to 250 customers.**

   - **Given data is good data or bad data ?**

     Any data cleansing operation is not required as there is no bad/dirty data present
   - **Select regression type as per type of target variable**

     As target variable is numeric**, linear regression is good to predict**
   - **What are predictor variable**

     By looking into p_values of predictor variable and scatter plot with target variable, we concluded that customer-segment and Avg_Num_Products_Purchased are predictor variable
   - **Select model if R-sqr value is greater than 0.7?**

     As adjust r-square value is **0.8366** which is greater than 0.7. So we **consider our model is good predictive model** as per predictive variable for target variable

2. What data is needed to inform those decisions?

   We would need a customer segment data i,e customer must be included into below segment as it is predictor variable.

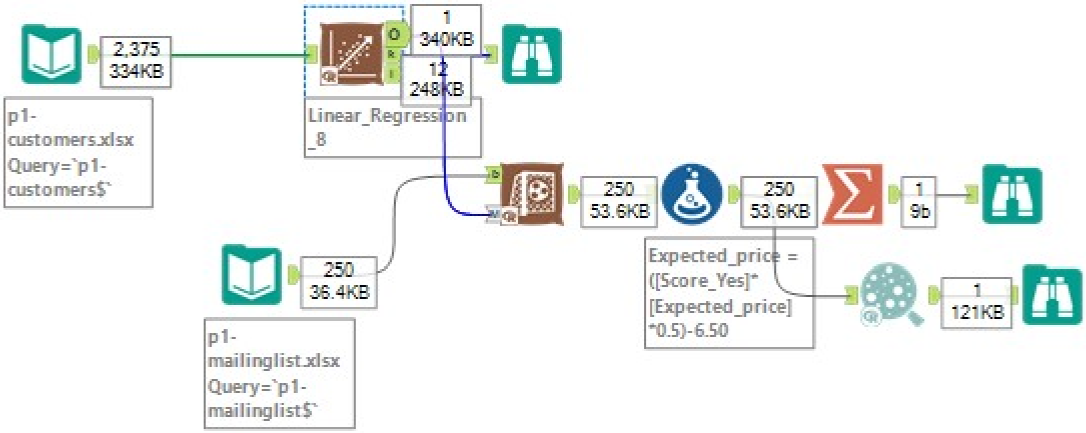
   Credit Card Only, Loyalty Club Only, Loyalty Club and Credit Card, Store Mailing List

   Along with customer segment data, we would need average of product purchased by customer.

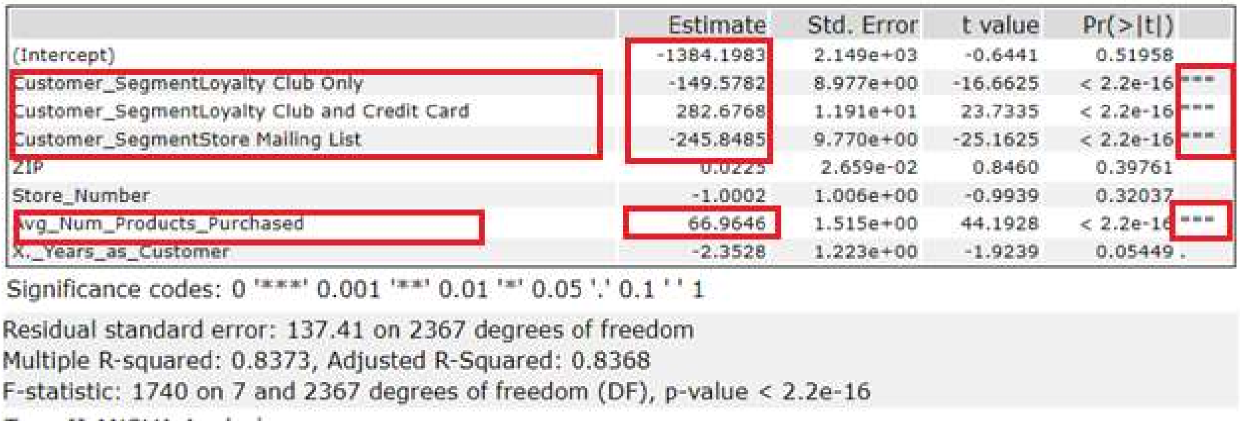
Step 2: Analysis, Modeling, and Validation

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.


2,375
334KB
p1-customers.xlsx Query=`p1-customers$`
1
340KB
12
248KB
Linear_Regression_8
250
53.6KB
250
53.6KB
1
9b
Expected_price = ([Score_Yes]*[Expected_price]*0.5)-6.50
1
121KB
250
36.4KB
p1-mailinglist.xlsx Query=`p1-mailinglist$`

In this model, avg_sale_amount varies as per zip_code, store_number, Avg_Num_Products_Purchased, #_year_as_customer. customer_segment Please find below relation between them.

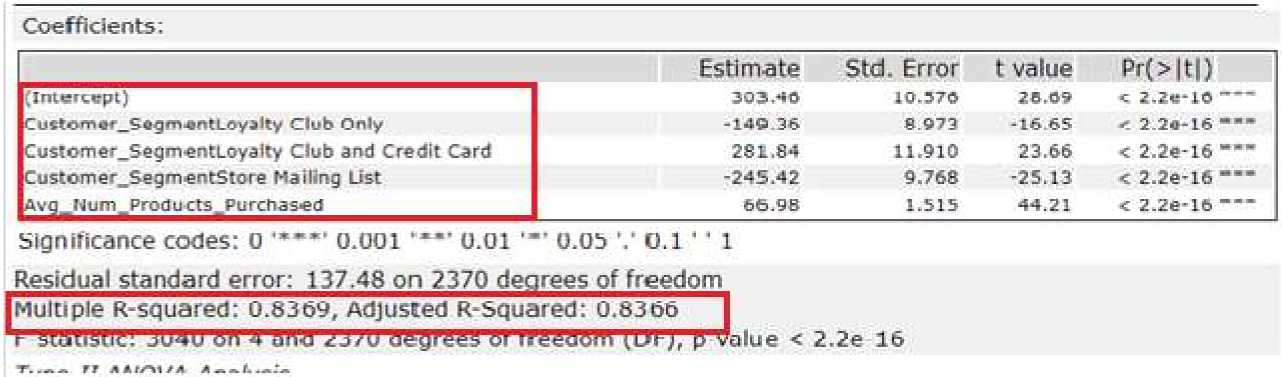


	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958	
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00	-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

Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

But we can see that, there is strong relationship between **Target variable(Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased , customer_segment category) having p-value and R-squared as per above img. There is NO significance between target variable and predictor variable zip code, store_num, #_year_as_customer. So we consider not to include them in equation.** Here we consider **Target variable (Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased, customer_segment).**



Coefficients:

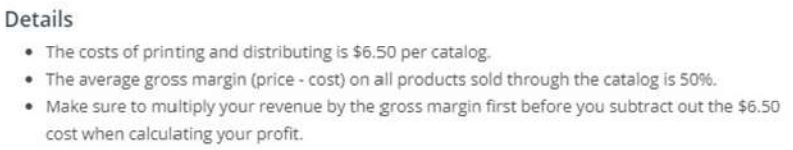
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303.46	10.576	28.69	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16	***
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

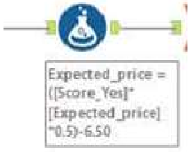
As adjust r-square value is **0.8366** which is greater than 0.7. So we consider our model is good predictive model as per predictive variable for target variable

Below logic is implemented as per given



Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Make sure to multiply your revenue by the gross margin first before you subtract out the $6.50 cost when calculating your profit.



Expected_price = ([Score_Yes]*[Expected_price]*0.5)-6.50

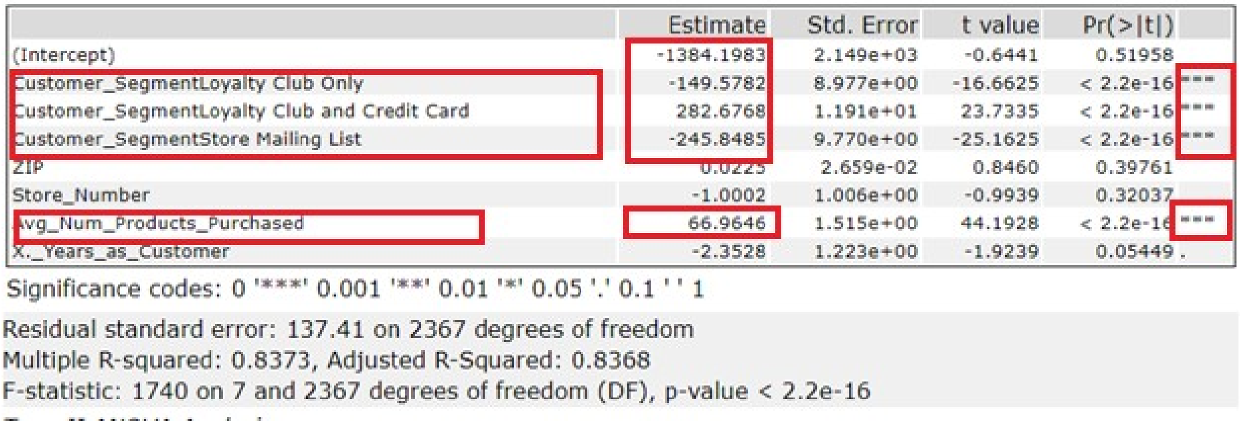
So we got expected profit 21987.43 Company should send catalog to 250 customers.

1. **How and why did you select the predictor variables in your model? You mu t 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.

By looking into data, it is been observed that Avg_sale_amount varies as per zip code, store_num , Avg_num_of_product_purchased, #_year_as_customer,customer segment

In this model, Please find below relation between them.

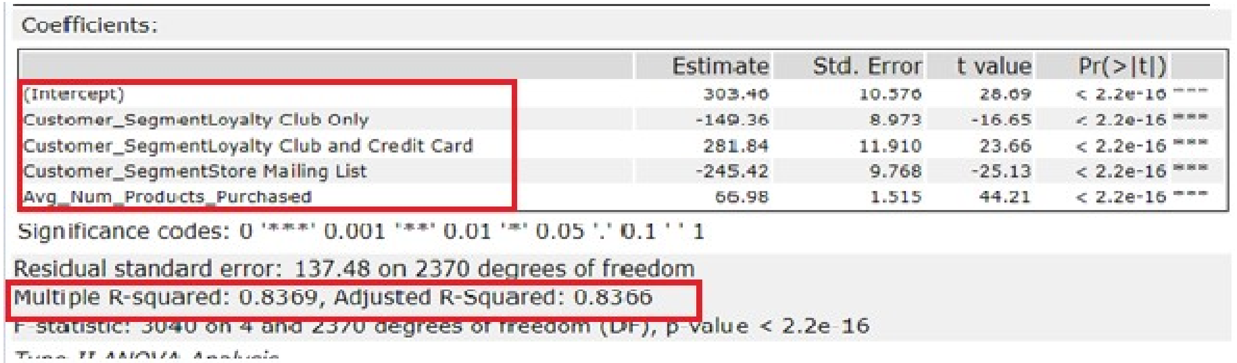


	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1384.1983	2.149e+03	-0.6441	0.51958	
Customer_SegmentLoyalty Club Only	-149.5782	8.977e+00	-16.6625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.6768	1.191e+01	23.7335	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.8485	9.770e+00	-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

Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

we can see that, there is strong relationship between **Target variable(Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased , customer_segment category) having p-value and R-squared as per above img. There is NO significance between target variable and predictor variable zip code, store_num, #_year_as_customer. So we consider not to include them in equation.** So I have decided to consider variable for those having strong relationship as per p_value. Here we consider **Target variable (Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased, customer_segment).**



Coefficients:

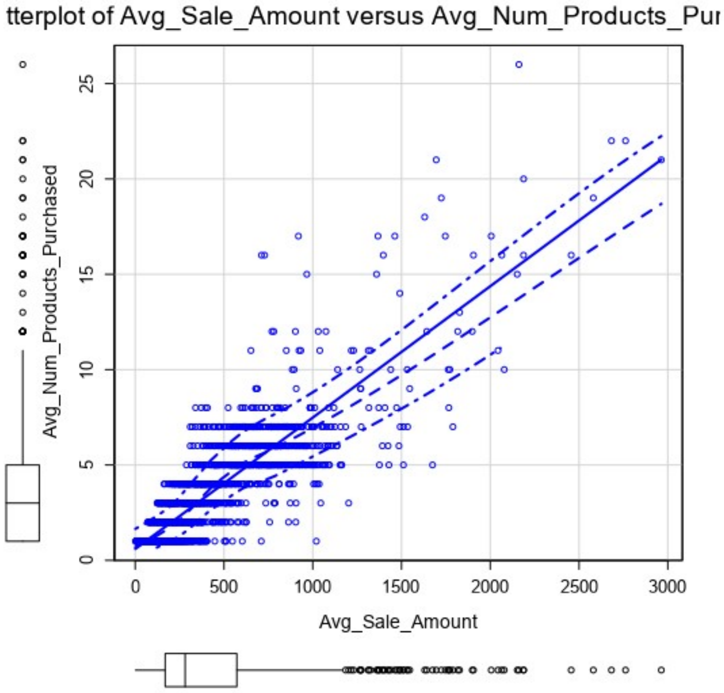
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303.46	10.576	28.69	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16	***
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

As you can see in above ,p-value for categorical variable is less than 0.05. it is observed that there is relationship between customer_segment category and avg_sale_price. There is significance for customer_segment. Hence we consider it is as predictor variable. Same case is applied for Avg_Num_Products_Purchased, it is observed that there is relationship between Avg_Num_Products_Purchased and avg_sale_price.

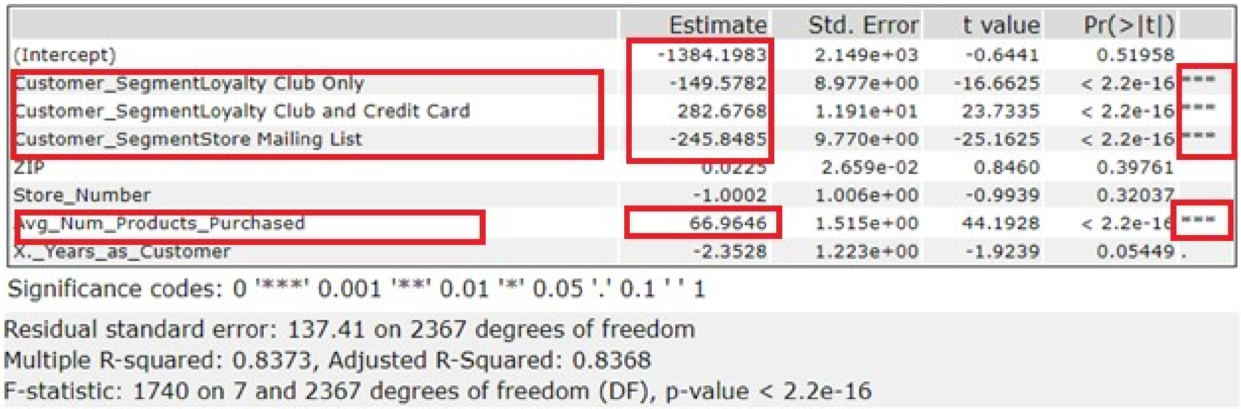
As you see in below graph, there is linear relationship between Target variable (Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased)



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.

With the help of linear model, we can find out the relationship between target variable and predictor variable. We can also find strong relation between variable with help of p-values and R-squared values.

First after looking into data, it is been observed that Avg_sale_amount varies as per zip code, store_num , Avg_num_of_product_purchased, #_year_as_customer.



	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

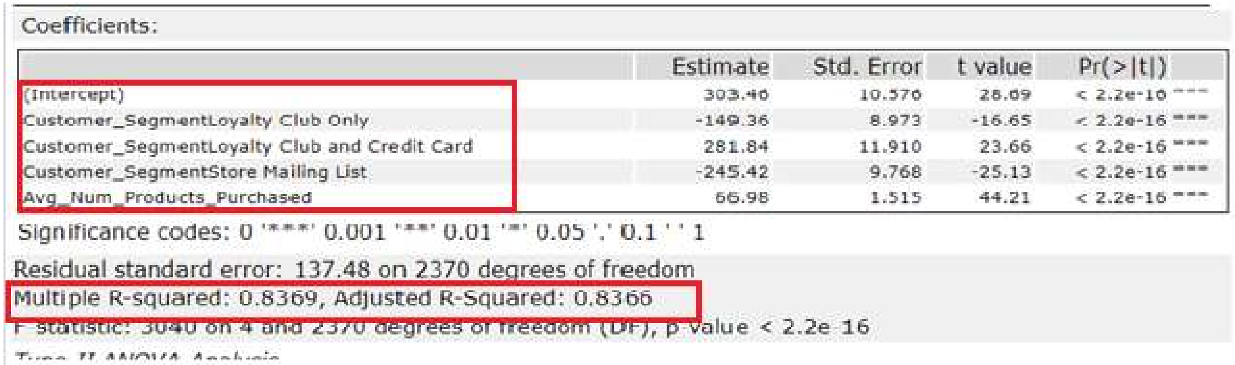
Residual standard error: 137.41 on 2367 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.8368
F-statistic: 1740 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16

As per above, there is strong relationship between **Target variable(Avg_sale_price) and predictor variable(Avg_Num_Products_Purchased, customer_segment) based on p-value. There is NO significance between target variable and predictor variable zip code, store_num, #_year_as_customer. So we consider not to include them in equation**

.Here we consider **Target variable (Avg_sale_price) and predictor**

variable(Avg_Num_Products_Purchased, customer_segment).

.



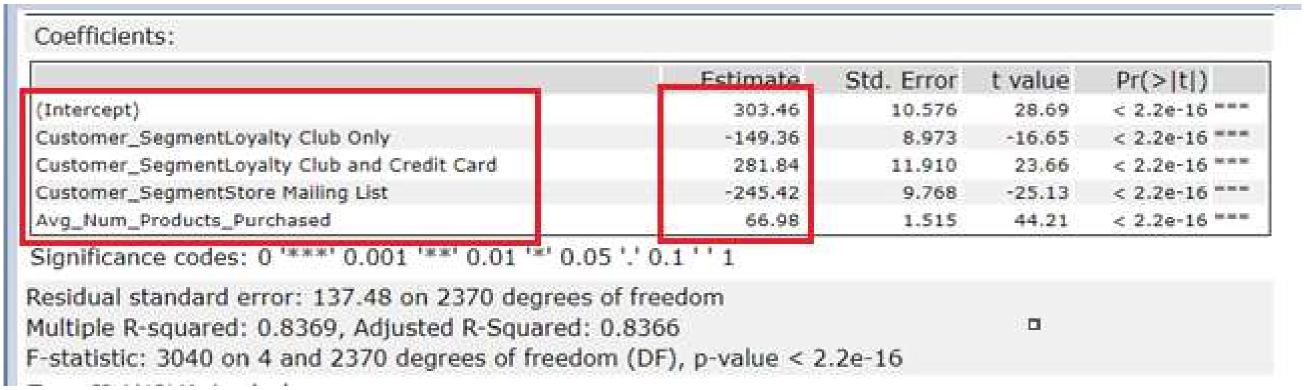
```
Coefficients:
                                            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                                   303.46      10.576    28.69  < 2.2e-16 ***
Customer_SegmentLoyalty Club Only            -149.36       8.973   -16.65  < 2.2e-16 ***
Customer_SegmentLoyalty 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
```

As per above, p-value for categorical variable is less than 0.05. it is observed that there is relationship between customer_segment category and target variable. There is significance for customer_segment. Hence we consider it is as predictor variable. Same case is applied for Avg_Num_Products_Purchased, it is observed that there is relationship between Avg_Num_Products_Purchased and target variable.

As **adjust r-square** value is **0.8366** which is greater than 0.7. So we consider our model is good predictive model as per considered predictive variable for 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)

Expected_sale_price =33.46+ Avg_Num_Products_Purchased*66.98+ Customer_segmentLoyalty Club Only*(-149.36)+ Customer_segmentLoyalty Club and Credit Card*(281.84)+ Customer_segmentStore Mailing List*(-245.42)

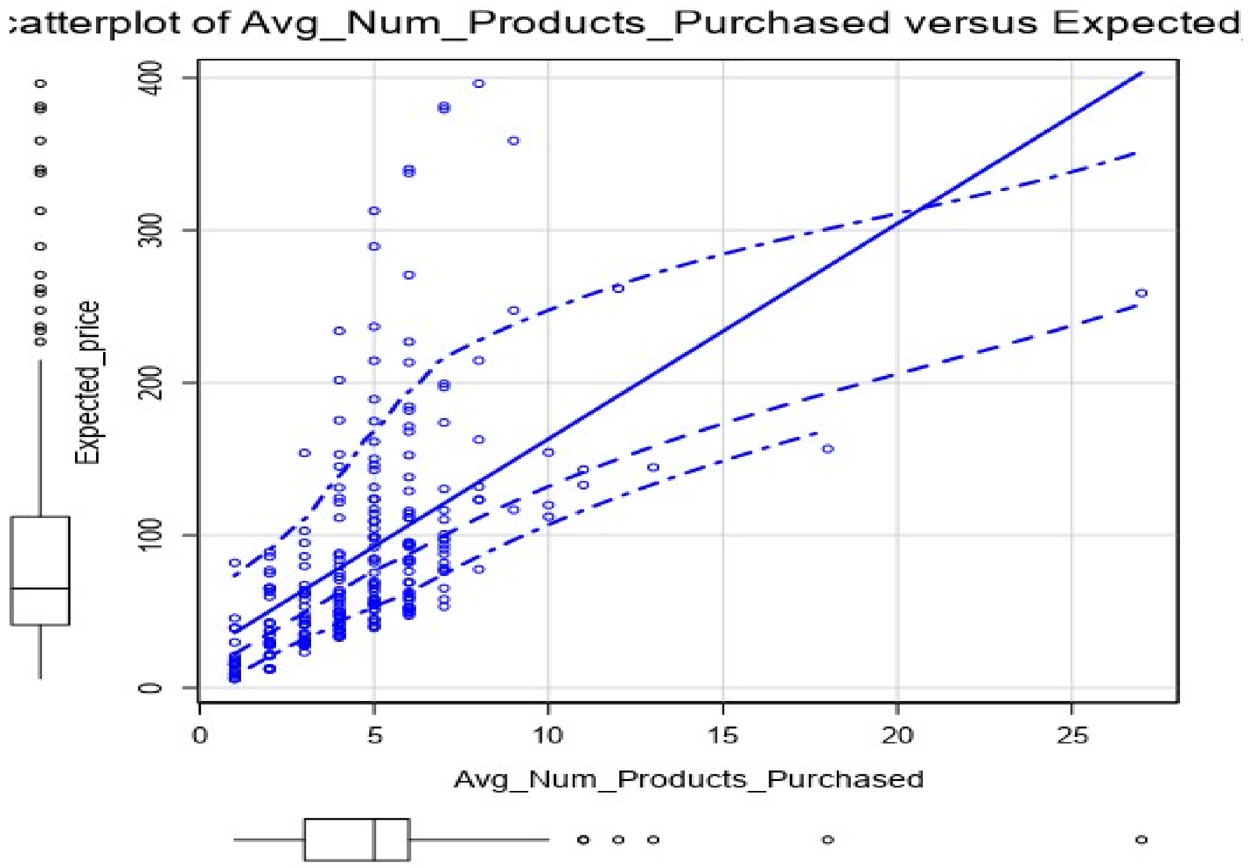


```
Coefficients:
                                            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                                   303.46      10.576    28.69  < 2.2e-16 ***
Customer_SegmentLoyalty Club Only            -149.36       8.973   -16.65  < 2.2e-16 ***
Customer_SegmentLoyalty 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
```

Step 3: Presentation/Visualization

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

Please find below scatter plot graph by considering new data with predicted model.



At the minimum, answer these questions:

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

As per above graph, we can see that Expected_profit than $10,000. So, Company should send catalog to 250 customers.

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

As per my analysis, sum of expected profit is 21987.43. Here, expected profit is calculated by using formula regression formula

Expected_sale_price =33.46+ Avg_Num_Products_Purchased*66.98+ Customer_segmentLoyalty Club Only*(-149.36)+ Customer_segmentLoyalty Club and Credit Card*(281.84)+ Customer_segmentStore Mailing List*(-245.42)

We calculated final price **([Score_Yes]*[Expected_price]*0.5)-6.50** which is as per below logic.

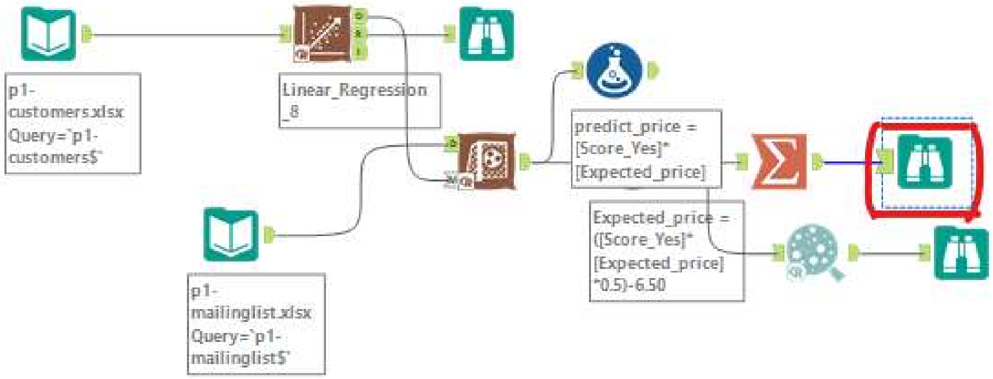
Details

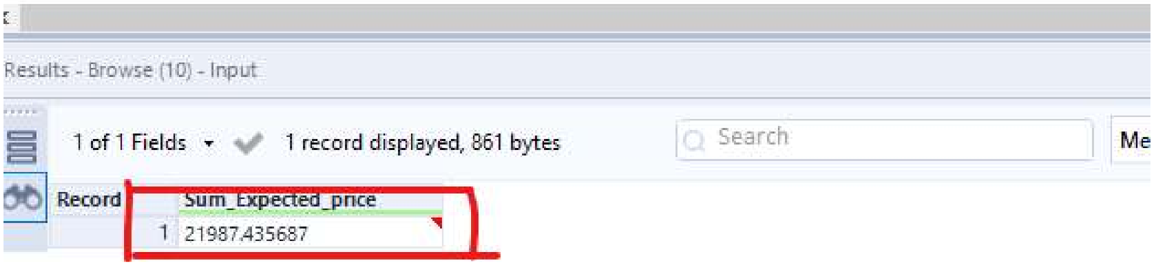
- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Make sure to multiply your revenue by the gross margin first before you subtract out the $6.50 cost when calculating your profit.

Where is expected price is generated with help of model. Finally, we take sum of all expected_price , we got 21987.43.

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

21987.43.





Results - Browse (10) - Input

1 of 1 Fields ▾ ✓ 1 record displayed, 861 bytes Search

Record	Sum_Expected_price
1	21987.435687