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)

Business understanding is observed as there are some retail store and they sell some items. They also record their customer activity. Dataset is of customer activity and their activity aggregated at customer level. The data set contains 12 features with avg_sales as target feature and no missing values.

Key Decisions:

1. What decisions needs to be made?
   Ans. Decisions is to find potential sales if the new catalog of items were introduced to the listed customers and by using likelihood score per customer to estimate sales value.

2. What data is needed to inform those decisions?
   Ans. Historical data of the customers with their purchase and interest in our store. This data allows us the estimate the sales given the catalog.

Step 2: Analysis, Modeling, and Validation

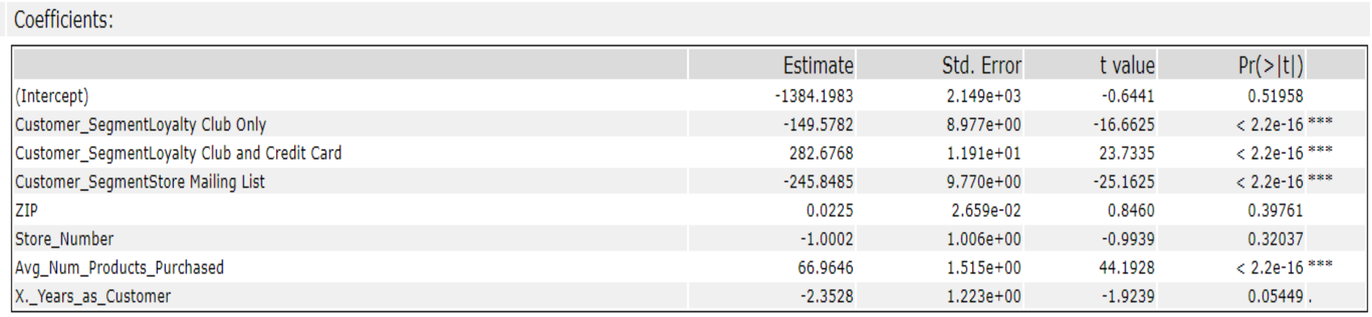
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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.

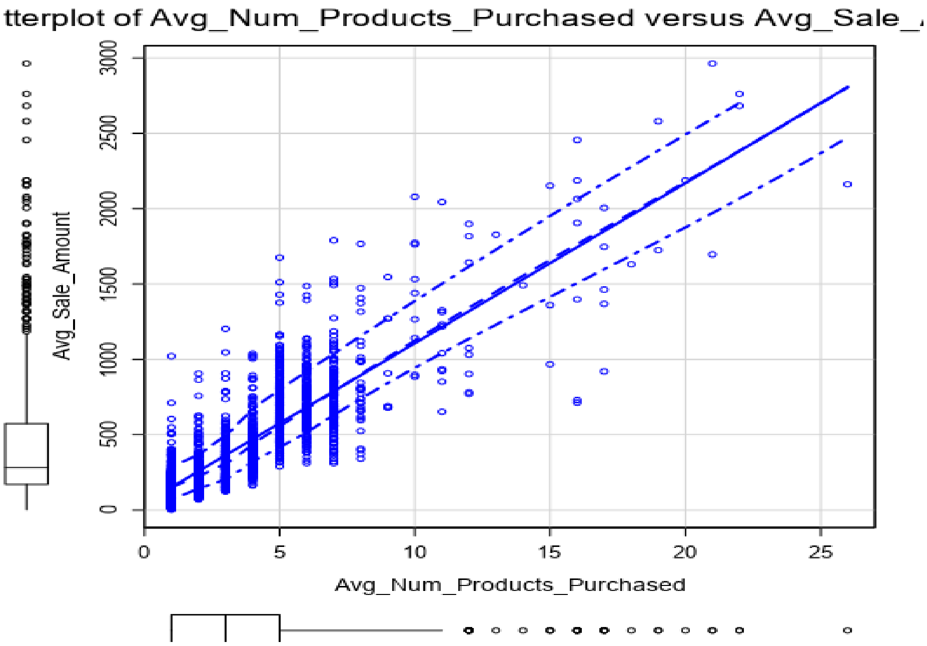
Ans.

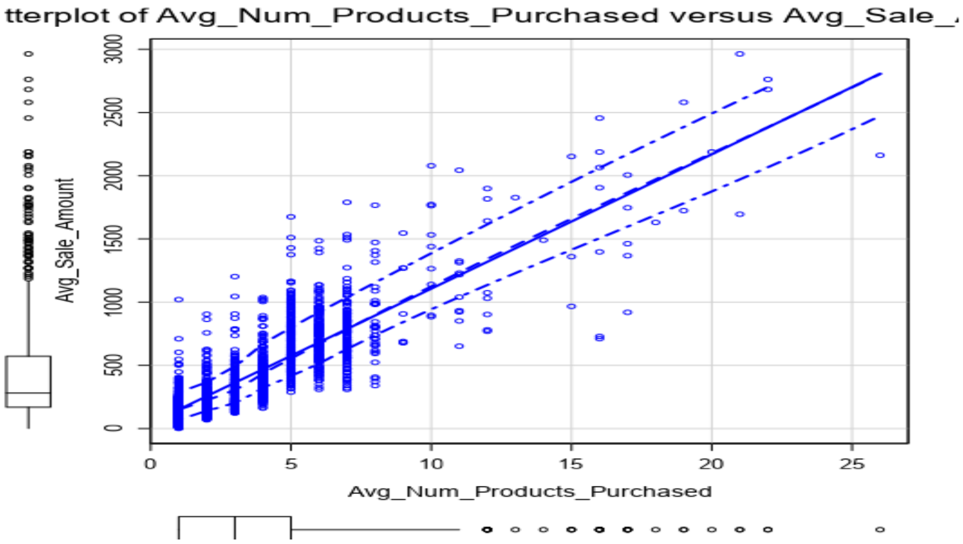


Coefficients:

	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 .

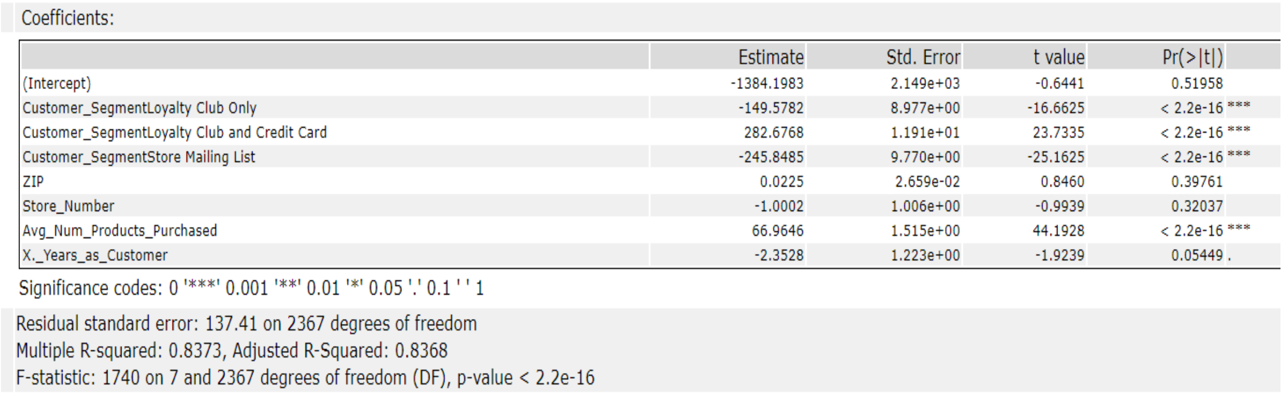
From the above report we can find out that only 2 variables are <= 0.05 p-value in other word statistically significant.





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.

Ans.

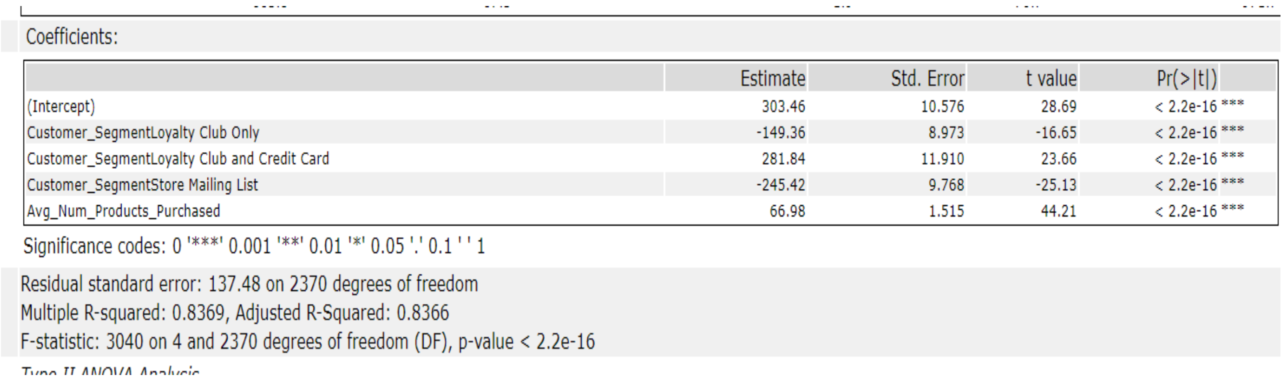


Coefficients:

	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



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

"In **statistics**, the **p**-**value** is the probability of obtaining results as extreme as the observed results of a **statistical** hypothesis test, assuming that the null hypothesis is correct. ... A smaller **p**-**value** means that there is stronger evidence in favor of the alternative hypothesis"

__ Google

As, p -value have that significance in regression then I tried to use that p-value as parameter to remove a feature and see if I can improve previous accuracy (R2_score value) but it is seen that by removing the feature we able to improve that significate but a insight emerges that the feature which are removed are contributing anything much.

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)

Ans. Est Sales Value = ***303.46*** *- 149.36 ** ***Customer_SegmentLoyalty Club Only*** *+ 281.84 ** ***Customer_SegmentLoyalty Club and Credit Card*** *- 245.42 ** ***Customer_SegmentStore 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?

Ans. Yes I recommend to introduce our new catalog to the customers

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

Ans.

bin	**0-20**	**20-40**	**40-60**	**60-80**	**80-100**
Avg	37.89186	69.89235	128.2626	182.4287	242.54

Bin Score Yes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
20-40	150	60.00	150	60.00
0-20	40	16.00	190	76.00
40-60	31	12.40	221	88.40
60-80	17	6.80	238	95.20
80-100	12	4.80	250	100.00

Bin Score Yes	Frequency	Acquired %	Avg net profit per cust.	total
20-40	150	70.00	88.00	9240
0-20	40	10.00	88.00	325
40-60	31	80.00	88.00	2200
60-80	17	90.00	88.00	1408
80-100	12	100.00	88.00	1056
total	250	66.00	88.00	14229

It is fair to assume that there is very less probability that customer from bin 0-20 will buy from over catalog but will recommend to introduce to them also. From above insight I have taken global avg to calculate expected profit which inherently have 72.40 per cust. as std. deviations due to high ticket size for high probability bins.

Note that the Acquired % is a assumption due to lack of historical data and insights.

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

Ans. The profit estimation are 14,229 to 21,987 if we introduce our catalog to 250 customers which almost 30 to 50% of over investment

Metric	**Value**
Sum_Predicted_Sales	138292.1301
Sum_Sales_yes_value	47224.87134
Sum_average gross margin	23612.43567
Sum_net profit	21987.43567
Count	250
Median_net profit	64.98418427
Avg_net profit	87.94974269
StdDev_net profit	72.39440262

.