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?

    - *The manager must decide if it is profitable to send out the catalog to the 250 new customers. Therefore, the manager set a threshold of 10,000 $ additional net profit produced by sending the catalog to the new customers.*

2. What data is needed to inform those decisions?

    - *To determine the additional profit, we need to predict the average amount of sales for each new customer and the probability of them purchasing products. We must consider the planned gross margin as well as the costs for printing and sending the catalogs to the new customers. The predicted net profit will be the data to inform the decision.*
    - *To predict the average amount of sales we need historical data of customers and their amount of sales.*
    - *The prediction model is then used on the dataset of new customers to create the predicted average amount of sales.*

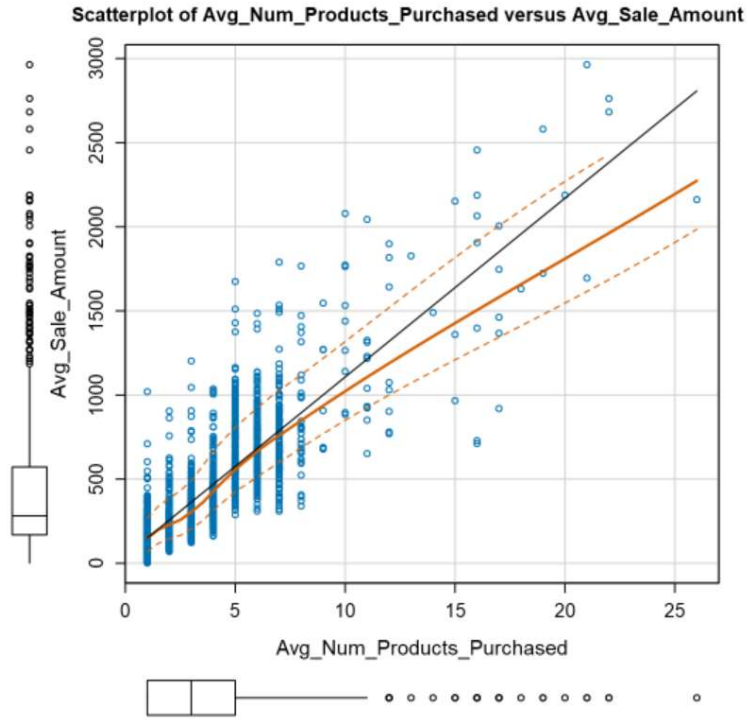
Step 2: Analysis, Modeling, and Validation

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)

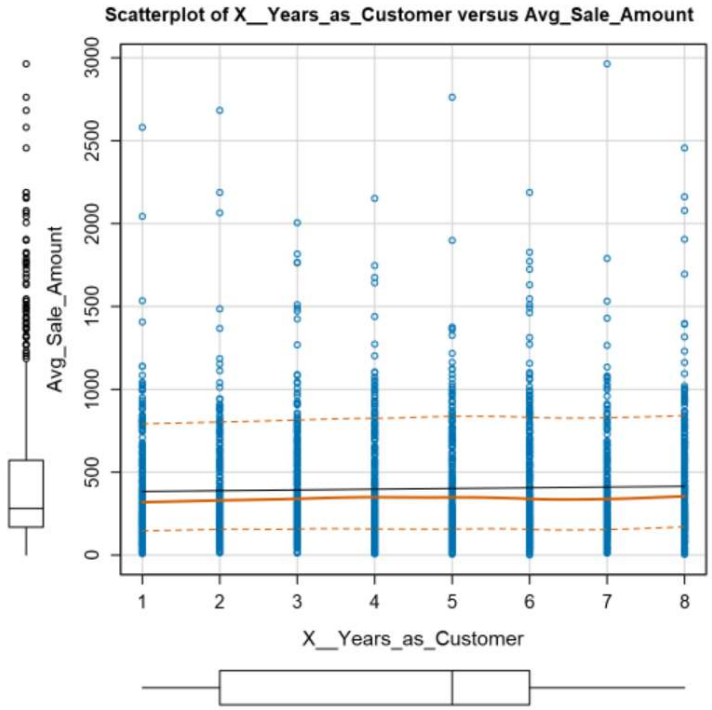
The first step was to look at the complete dataset of p1-customers.xlsx. Some of the variables for each customer are obviously individual to each customer and have no relation to the average amount of sales. For example, Customer Names, the customer ID or the Address. The variable "State" has only one data point, so we can exclude this variable. In the Next step it made sense to classify each variable as a categorial or numerical variable. After reviewing each variable, the following classification was derived:

- *Categorial variables: Customer Segment, City, ZIP, Store Number, Response to last catalog.*
- *Numerical: Average number of products purchased, Years as a customer.*

The average number of products purchased has a linear relationship to the target prediction variable. Therefore, it was included as a prediction variable.



The variable Years as a customer, has no visible relationship to the target variable. Therefore, it was excluded from the model as it would not make a good prediction variable. Another reason is that the data set which will be used to make predictions has only customers with a time frame of 1 year and below.



After using trial and error with every categorial variable we have the following solution:

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	***

This suggest a good statistical significance for the variable.

City

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	46.997	11.968	3.92678	9e-05	***
CityAurora	-14.529	13.726	-1.05851	0.28993	
CityBoulder	-145.347	102.297	-1.42084	0.1555	
CityBrighton	-111.702	124.973	-0.89381	0.37152	
CityBroomfield	4.241	19.319	0.21953	0.82626	
CityCastle Pines	-120.845	124.957	-0.96710	0.33359	
CityCentennial	8.732	22.861	0.38198	0.70251	
CityCommerce City	-2.722	56.894	-0.04784	0.96185	
CityDenver	1.193	12.913	0.09241	0.92638	
CityEdgewater	34.615	52.029	0.66530	0.50593	
CityEnglewood	11.914	26.050	0.45736	0.64745	
CityGolden	-37.709	41.902	-0.89993	0.36825	
CityGreenwood Village	-116.235	48.358	-2.40366	0.01631	*
CityHenderson	-151.877	176.356	-0.86120	0.38922	
CityHighlands Ranch	24.724	38.370	0.64436	0.5194	
CityLafayette	-107.274	79.506	-1.34925	0.17739	
CityLakewood	9.785	16.378	0.59745	0.55026	
CityLittleton	-22.425	23.572	-0.95135	0.34153	
CityLone Tree	62.910	176.428	0.35658	0.72144	
CityLouisville	-71.011	88.710	-0.80048	0.42351	
CityMorrison	40.004	67.466	0.59295	0.55327	
CityNorthglenn	-49.421	37.631	-1.31330	0.18921	
CityParker	-23.356	35.676	-0.65468	0.51274	
CitySuperior	-84.896	59.726	-1.42144	0.15532	
CityThornton	24.978	31.788	0.78576	0.43209	
CityWestminster	-8.311	22.133	-0.37550	0.70732	
CityWheat Ridge	18.366	26.441	0.69460	0.48738	
Avg_Num_Products_Purchased	106.423	1.325	80.29742	< 2.2e-16	***

This suggests that the variable has with one exemption no statistical significance.

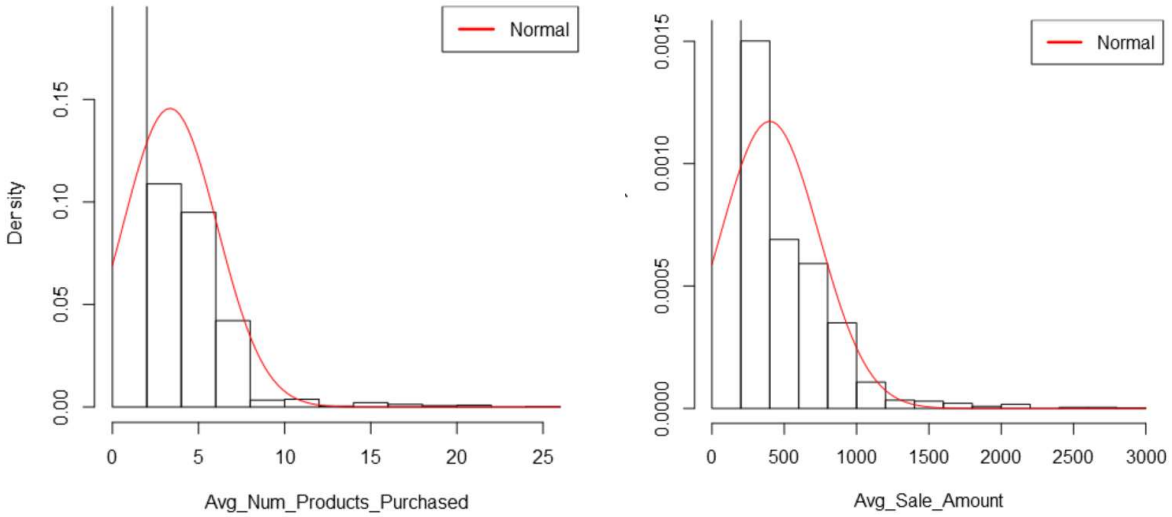
the same process was used with "ZIP" and "Store Number" which showed no statistical significance. In addition, the variable "response to catalog" is only available in one dataset so it can't be used to develop the model. This is unfortunate as it seems to have some statistical significance.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	305.00	10.582	28.823	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16	***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241	*
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16	***

The final prediction variables are the average number of products purchased and customer segment. As these two are the only usable variables with a p-value suggesting a good statistical significance.

If we look at the distribution of the data "average sale amount" & "average number of products purchased":



We see that the distribution is skewed and not normally distributed. This suggests that we should gather more data or talk with our superior to get a better dataset. For now, it is assumed that the dataset represents all data available.

The Regression result for the model are as follows:

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	***

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 DF, p-value: < 2.2e-16

The R-squared suggest that around 83,69% of the variance in the target variable can be explained by our chosen prediction variables and the remaining 16,31% of the variance remains unexplained by our model.
As we try to predict the human behavior of buying or not buying, I expected a much lower r-squared, this suggests that we have strong and usable model for the cause.

The regression equation is:

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

Step 3: Presentation/Visualization

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

The company should send the catalog to the 250 new customers based on the results of the liner regression.

The question identified, was if the catalog will provide more than 10,000 $ in additional net profit if we send it to the 250 new customers. Based on this question the provided data was analyzed and the variable “average sale amount” was identified as the target variable to be predicted. After testing the remaining variables on their linear relationship and statistical significance, the variables “average number of products purchased” and “customer segment” were chosen as prediction variables. The results of the model suggested that it will be a good fit to predict the target variable. This were derived by the low p-values.

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	***

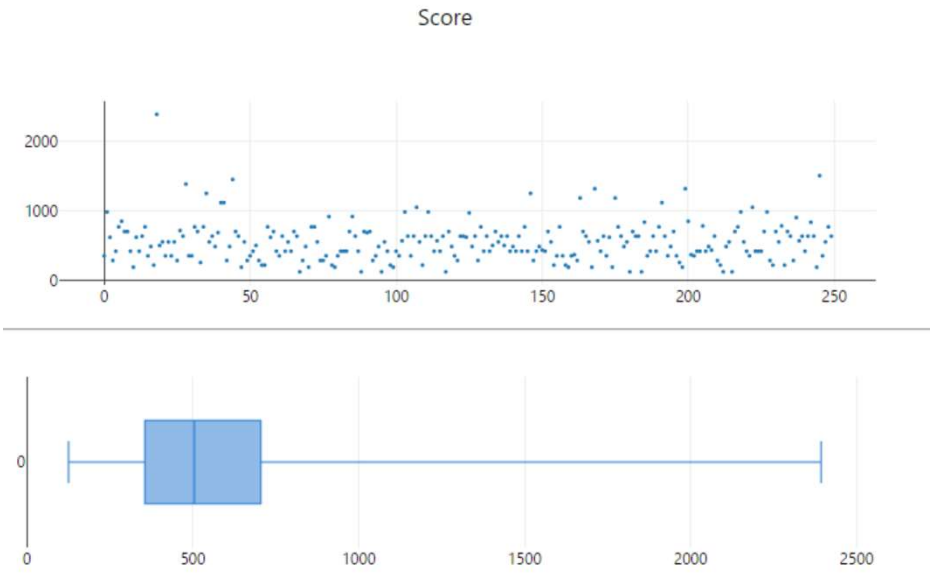
And the r-squared of 83,69%

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 DF, p-value: < 2.2e-16

The model was then used on the dataset of the 250 new customers to predict their average sale amount. The equation of the model can be expressed as:

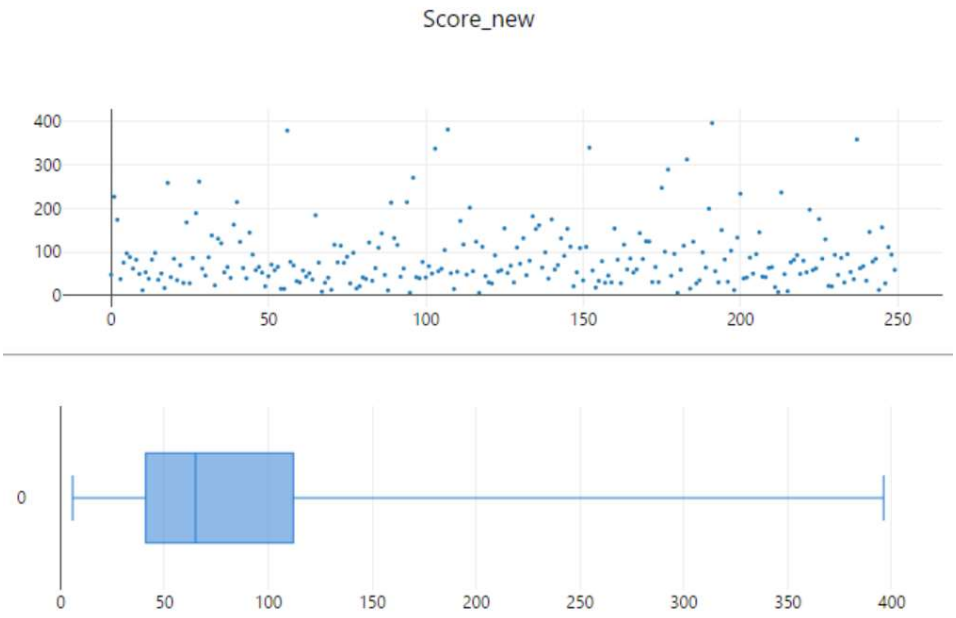
*Y= 303.46 + 0 (If Type: Credit Card Only) -149.36 (If Type: Loyalty Club Only) + 281.84 (If type: Loyalty Club and Credit Card) – 245.42 (If Type: Store Mailing List) + 66.98 * Avg_Num_Products_Purchased*

Results of the prediction model:



Then the expected probability of each new customer was multiplied with the results of the prediction model. The gross margin was accounted and the costs for sending and printing the catalog to each customer was subtracted.

Results after including probability, gross margin and costs:



After summing up all results an additional net profit of 21,987.44 $ in total is expected. Derived from this result, we can suggest sending out the catalog to the 250 new customers.