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?**

   Whether the company will send the catalogs out to 250 new customers from the company's mailing list.

2. **What data is needed to inform those decisions?**

   To inform the decision of the company sending out catalogs we need to calculate:

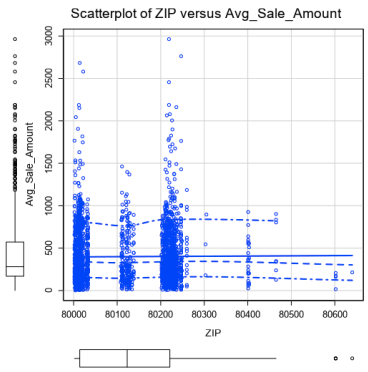
   - ***'predicted average sales amount'** for the new customers;*
   - ***'expected sale'** (taking into account the probability of purchase);*
   - ***'expected profit'** (taking into account the margin and the cost of sending the catalogs out).*
   - *To compare 'expected profit' with $10,000 (minimal expected profit by management).*

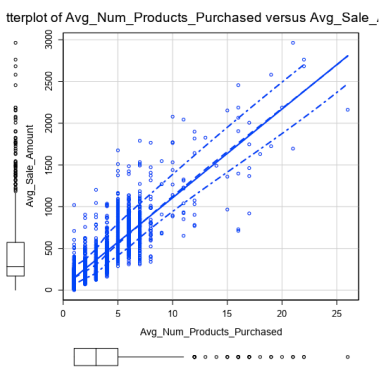
Step 2: Analysis, Modeling, and Validation

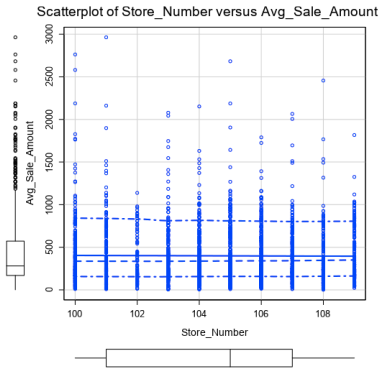
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

To predict profit for the new customers we need to predict 'average sale amount' for the 250 new customers from the company's mailing list. It means that '**average sale amount'** is the **target variable** for the linear regression model.

Understanding relations between each predictor variable and the target variable. To identify whether a linear relationship exists for categorical variables we should build scatterplots between the 'average sale amount' variable and predictor variables. I created scatterplots for all numeric predictor variables as we cannot use a scatterplot or any other graph to see whether a linear relationship exists for categorical variables (I will check the categorical variables through the regression model and see if the coefficients turn out to be significant with a high multiple-R-squared). Among the numeric variables there are 'Average sale amount', 'Average products purchased' and 'Store number'. The only variable with a clearly seen linear relationship between is the **'Average products purchased'**. As the 'Average products purchased' increases, the 'Average sale amount' increases too.







Clearly seen that **two variables have statistical significance** for the target variable (average sale amount) among them **'customer segment'** and **'average number of product purchased'**.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	21684.0811	9526.5619	2.27617	0.02293	*
Customer_SegmentLoyalty Club Only	-149.2453	9.0334	-16.52147	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	284.1845	11.9806	23.72033	< 2.2e-16	***
Customer_SegmentStore Mailing List	-244.5549	9.8552	-24.81469	< 2.2e-16	***
CityAurora	-13.7146	16.7262	-0.81995	0.41233	
CityBoulder	12.8477	88.4601	0.14524	0.88454	
CityBrighton	73.3021	120.8646	0.60648	0.54425	
CityBroomfield	-15.2951	16.8477	-0.90784	0.36405	
CityCastle Pines	-71.6956	100.6902	-0.71204	0.47651	
CityCentennial	7.6706	22.5335	0.34041	0.73358	
CityCommerce City	-29.6307	45.2203	-0.65525	0.51237	
CityDenver	61.5205	28.1507	2.1854	0.02896	*
CityEdgewater	93.4705	48.2538	1.93706	0.05286	.
CityEnglewood	37.7942	25.5323	1.48025	0.13894	
CityGolden	98.4425	57.2991	1.71804	0.08592	.
CityGreenwood Village	-18.1659	41.3567	-0.43925	0.66052	
CityHenderson	-104.2277	157.6923	-0.66096	0.50871	
CityHighlands Ranch	8.2583	34.8289	0.23711	0.81259	
CityLafayette	-50.8219	62.5881	-0.81201	0.41687	
CityLakewood	58.5459	29.6508	1.97452	0.04844	*
CityLittleton	7.1084	25.1083	0.28311	0.77712	
CityLone Tree	116.6975	139.3807	0.83726	0.40253	
CityLouisville	-43.2155	70.063	-0.61681	0.53742	
CityMorrison	111.8866	75.9795	1.47259	0.141	
CityNorthglenn	35.7303	40.4653	0.88299	0.37733	
CityParker	23.5479	37.1426	0.63398	0.52615	
CitySuperior	-65.9113	47.5359	-1.38656	0.16571	
CityThornton	87.5758	39.614	2.21073	0.02715	*
CityWestminster	-8.2679	18.1807	-0.45476	0.64932	
CityWheat Ridge	30.2134	22.6076	1.33643	0.18154	
ZIP	-0.2669	0.1191	-2.24214	0.02505	*
Store_Number101	-2.6219	12.1977	-0.21495	0.82982	
Store_Number102	6.2524	20.0416	0.31197	0.75509	
Store_Number103	0.8508	15.4809	0.05496	0.95618	
Store_Number104	-11.7265	13.1115	-0.89437	0.37122	
Store_Number105	-12.4169	12.0347	-1.03175	0.30229	
Store_Number106	-19.8632	12.4965	-1.5895	0.11208	
Store_Number107	-19.8987	13.1031	-1.51862	0.12899	
Store_Number108	-12.8227	14.9977	-0.85497	0.39265	
Store_Number109	12.8494	17.8663	0.7192	0.47209	
Avg_Num_Products_Purchased	67.1439	1.5281	43.93897	< 2.2e-16	***
X._Years_as_Customer	-2.5307	1.2365	-2.04663	0.04081	*

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

I have tried to run the model and have got a couple of errors in Alteryx and my results did not correspond to the answers in the quiz. After following the instructions given by a mentor I restrict the model to be run just with variables with statistical significance and I got following results:

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	11.91	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

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

The model is highly predictive as we have low P-values (< .00000000000000022) and a high R-squared (0.8369).

The best linear regression equation based on the available data is

Y (Predicted_Avg_Sale_Amount) = 303.46 – 149.36 * Customer_Segment_Loyalty Club only + 281.84 * Customer_Segment_Loyalty Club and Credit Card – 245.42 * Customer_Segment_Store Mailing list + 66.98 * Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

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

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

I recommend to send out the catalogues to the 250 new customers from the company's mailing list.

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

1) *I have got '**predicted average sales amount**' for the 250 new customers.*
2) *For each customer I have multiplied 'predicted average sale' by 'score yes' (probability of purchase) to generate the '**expected sale**' of each individual*
3) *I have multiplied the' expected sales' by the margin (0.5) and subtract the cost of sending the catalogs to arrive at the '**expected profit**' of each individual.*
4) *I have sum all 'expected profit' to calculate 'expected profit' from the new catalog which is **21,987.44***

I recommend to send out the catalogues because the expected profit is more than twice bigger than $10,000 which was minimal expected profit.

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

21,987.44