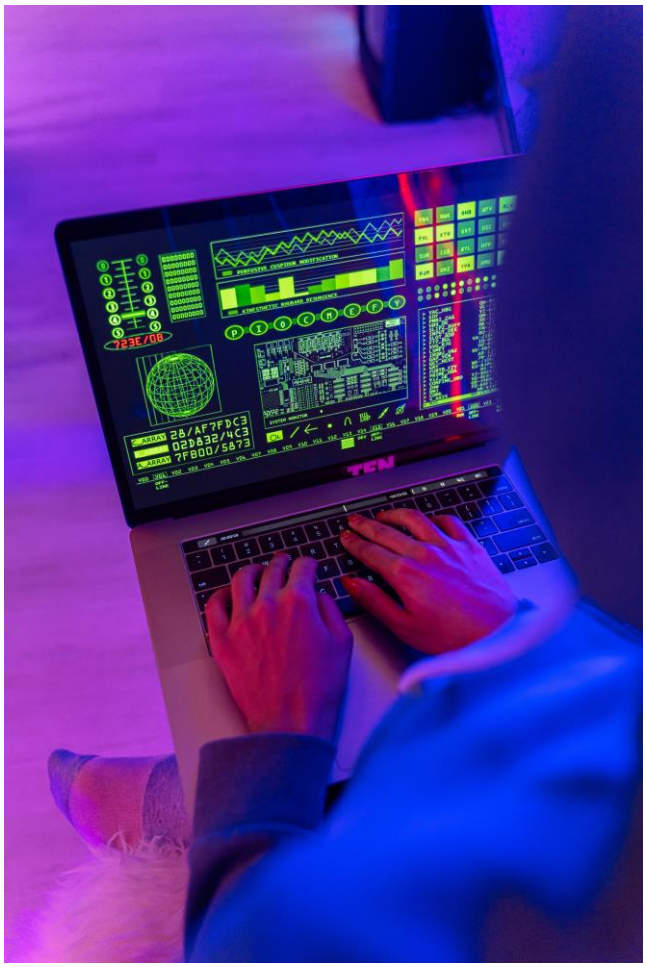

REPORT ON CATALOG DEMAND

A Predictive Analysis

ABSTRACT

This project is an analysis of a business problem in the mail-order catalog business. I've been tasked with predicting how much money your company can expect to earn from sending out a catalog to new customers. For this task, I have built a model and applied the results in order to provide a recommendation to management.

Oluwadewalade Ade-Onojobi

Predictive Analysis for Business Nanodegree

25/04/2021

LIST OF FIGURES

Figures **Pages**

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500-word limit)

Key Decisions:

1. **What decisions needs to be made?**

   The company I recently started working for sent out its first print catalog of high-end home goods they manufacture and sell last year and is preparing to send out this year's catalog to its 250 new customers in the coming months. These new customers were acquired from their mailing list. My manager has been asked to determine how much profit the company can expect from sending a catalog to these customers and I have been assigned to help your manager run the numbers as the business analyst. This is because my manager is knowledgeable about data analysis but not familiar predictive models.

   I have been asked to use my predictive modelling skills to build a model and predict the expected profit from these 250 new customers. I am to give a report on my analysis, a prediction, and a recommendation on whether to send the catalog. Management intends to make a profit of at least $10,000 from this catalog and will likely not want to send it out to these new customers unless my analysis guarantees them that much profit.

: :Awesome: Right! The main decision here is that the company wants to determine whether the expected profit from these customers exceeds $10,000 and then decide to send the catalog out to these customers or not.

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

   The decision to send the catalog or not depends on how much profit that would be expected to be made. To figure this out, we would need to train a predictive model using data the company has collected on past customers and make an informed prediction using that model. The data itself can be gotten from the company's database as they have previously collected data on past customers. The data would preferably be cleaned however I can clean the data if necessary. The data should include the number of purchases made by past customers, the number of years they have been a customer, the sources of these customers (customer segment), the total amount (in dollars of sales made to the customers, and their response to previous catalogues. This data would help me accurately predict the expected profit from the new customers.

: Suggestion: Note that we also need information on shipping and preparation costs to calculate profit.

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)

1. Analysis

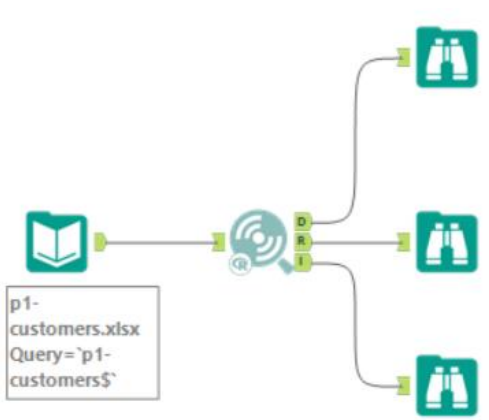

Fig 1: Workflow for analyzing categorical variables

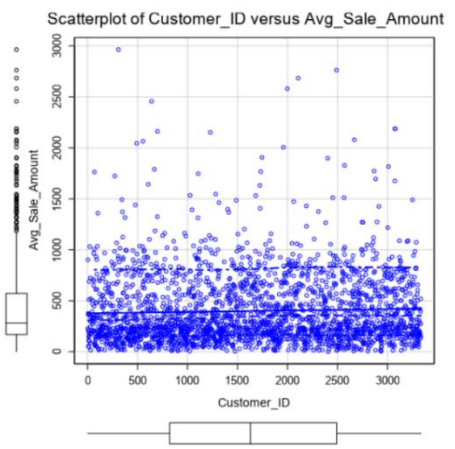

Fig 2: Scatterplot of Customer_ID vs. Avg_Sale_Amount

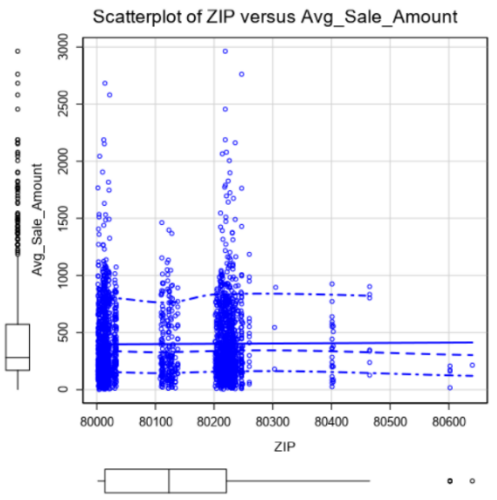

Fig 3: Scatterplot of ZIP vs. Avg_Sale_Amount

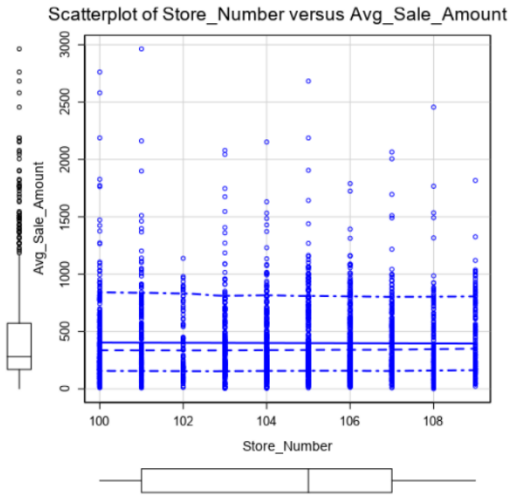

Fig 4: Scatterplot of Store_Number vs. Avg_Sale_Amount

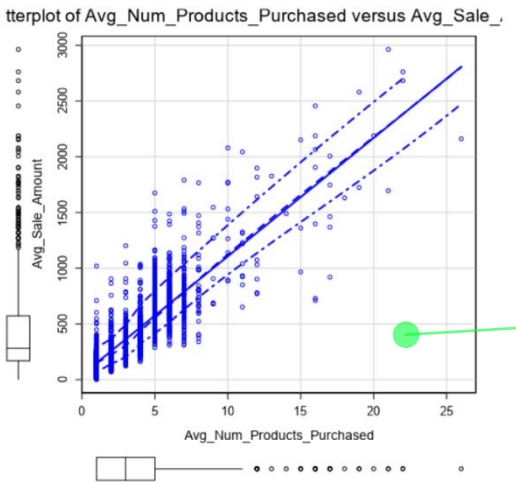

Fig 5: Scatterplot of Avg product purchased vs. Avg_Sale_Amount

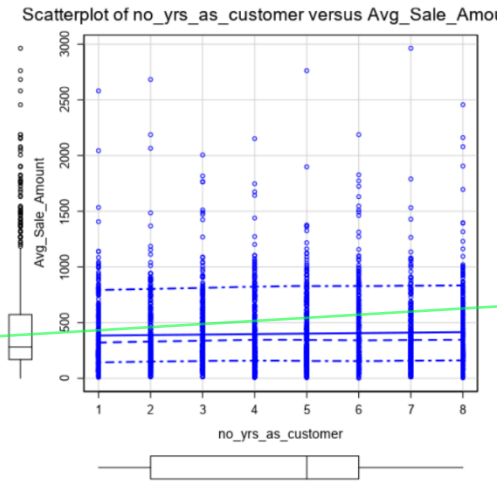

Fig 6: Scatterplot of Customer Years vs. Avg_Sale_Amount

: :Awesome: As this plot depicts, average number of products purchase is linearly related to average sale amount.

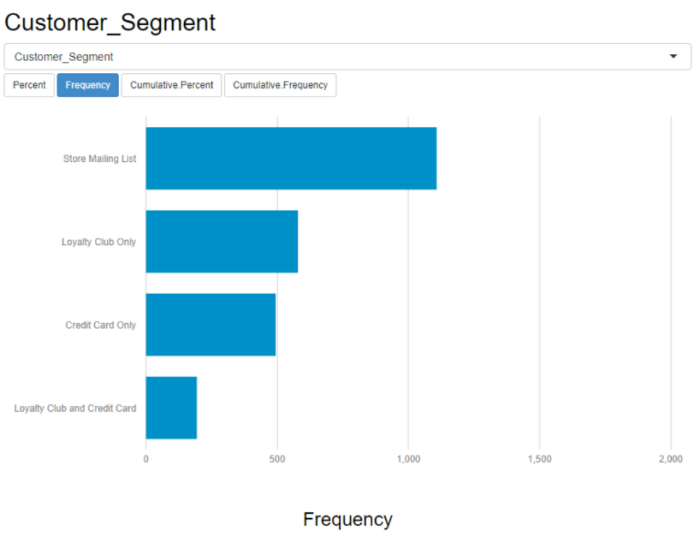

Fig 7: Frequency Distribution of Customer_Segment Categories

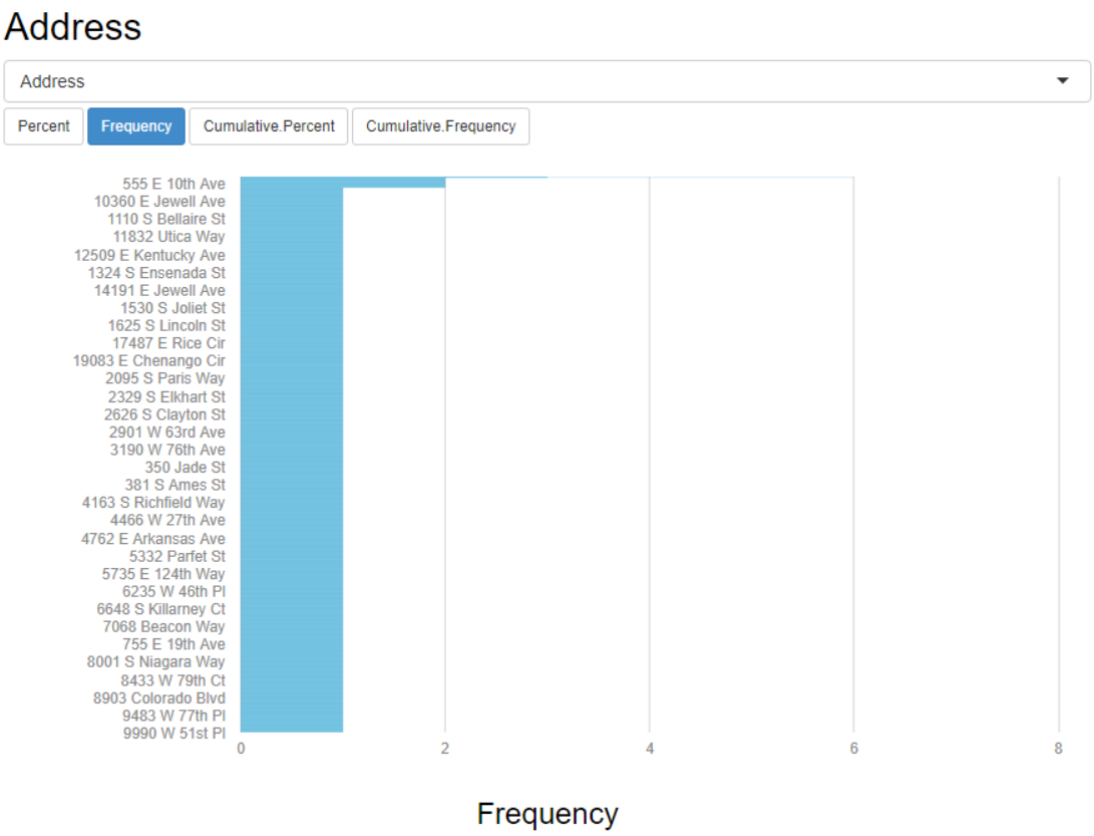

Fig 8: Frequency Distribution of Address Categories

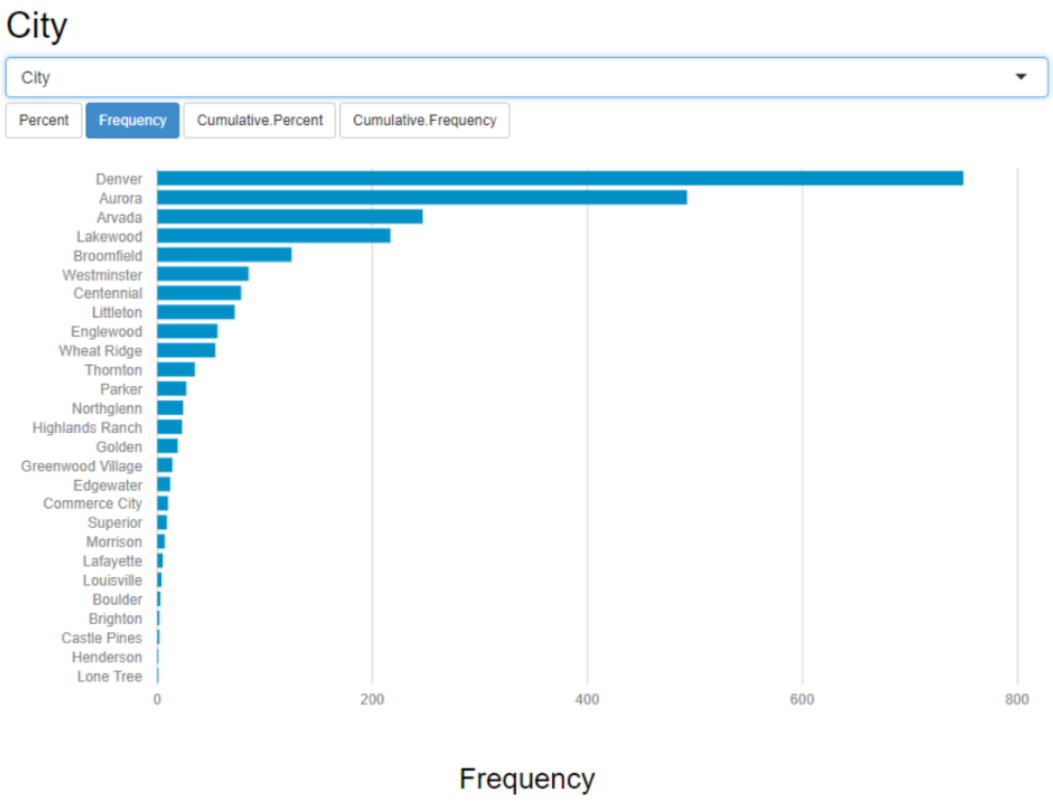

Fig 9: Frequency Distribution of City Categories

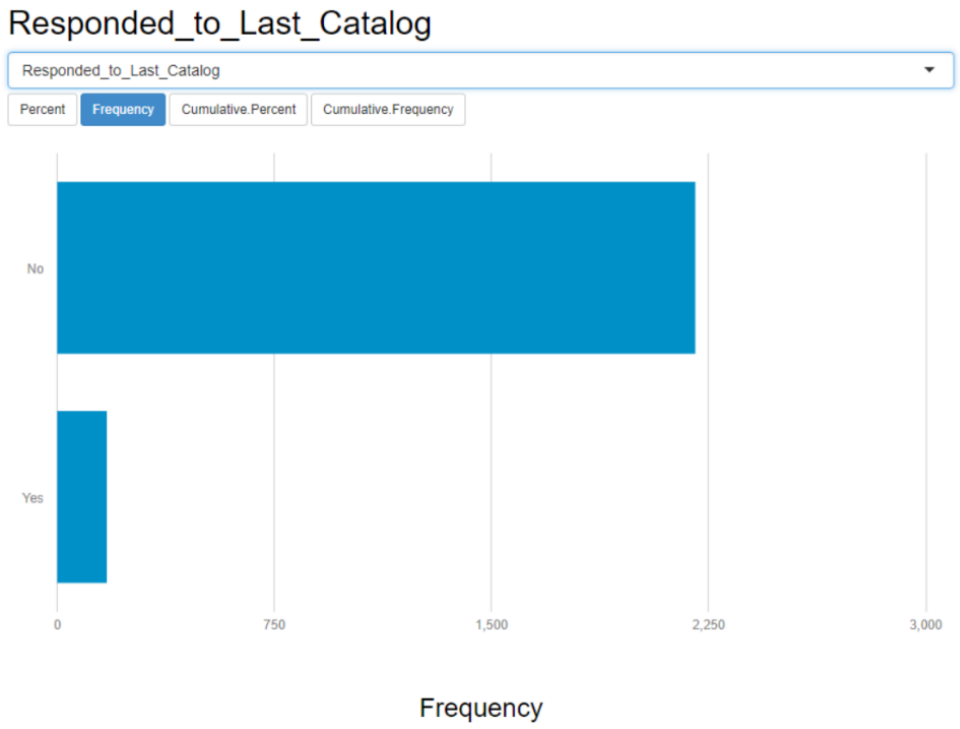

Fig 10: Frequency Distribution of Response_to_Last_Catalog Categories

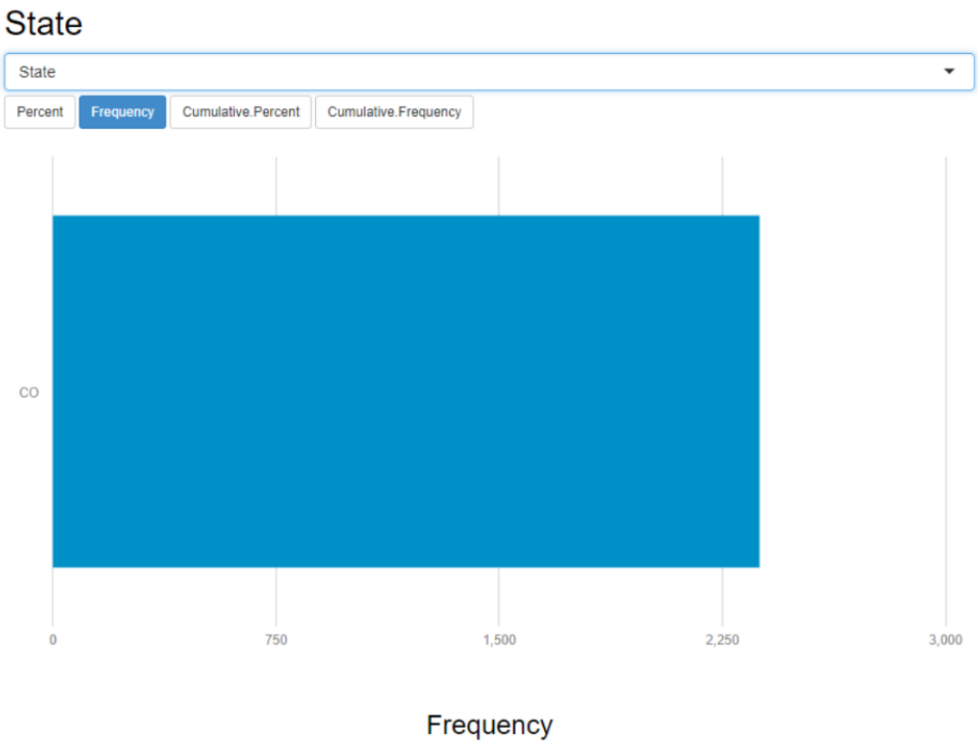

Fig 11: Frequency Distribution of Customer_Segment Categories

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 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.

To select my predictor variables, I analyzed variables from the p1_customers.xlsx file for correlation using the Avg_sale_amt variable as the target variable. Using the **scatterplot tool** in Alteryx, I checked for a linear relationship between the potential numeric predictor variables and Avg_sale_amt (the target variable). Figs 1 – 6 above show the scatterplots of all the numeric variables against the target variable. Only Avg_num_products_purchased (fig 5) showed a linear relationship with the target variable. We see an increase in Avg_sale_amt as the Avg_num_products_purchased variable increases.

Measuring linear relationship between categorical variables and the target variable using a scatterplot is hard as categorical variables can't be displayed as discrete points on a scatterplot. Because of this I performed a different analysis on the categorical variables. Looking at the figs 7 – 11, we notice they are bar charts not scatterplots. Fig 8 shows that most of the addresses are unique with a few appearing more than once and the highest appearance being six times. Fig 7 shows a steady decrease in the frequency of categories in the customer_segment variable with the highest number of them being **store mailing list** (1,108 customers) and the lowest being **loyalist club and credit card** (194 customers). From fig 9, a bulk of customers are from either Denver, Aurora, Arvada or Lakewood with very few being from areas between Superior and Lonetree on the chart. Fig 10 tells us that almost all customers on our list did not respond to the catalog sent the first time. Fig 11 shows that all customers come from the same state.

To conclude my analysis, I decided to use all the categorical variables in my model except the responded_to_last_catalog, address, and state variables. I discarded responded_to_last_catalog because this variable variable would not be useful as a predictor because it was not found in the p1_mailinglist.xlsx. I also removed the address because it would likely show little significance to predicting the target variable due to the number of unique values. Finally, the state variable was removed because the was only one category in this variable, so it won't have any effect in making a prediction. For the numerical variables, I chose the Avg_num_products_purchased variable because it showed a strong linear relationship with the target variable.

: Awesome: Good job explaining the process for selecting variables.

Due to the nature of categorical variables, we cannot use a scatterplot to see if there is a linear relationship to categorical variables. The best way to verify the linear relationship is to include the categorical variables in the regression model and see if the coefficients are significant and increase the adjusted r^2. If there is a linear relationship, then the coefficients are significant and the r^2 must be relatively high.

2. Modelling

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	22442.5757	1.092e+04	2.05482	0.04006	*
Customer_SegmentLoyalty Club Only	-164.8030	1.081e+01	-15.24814	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	239.2080	1.376e+01	17.38210	< 2.2e-16	***
Customer_SegmentStore Mailing List	-263.9644	1.188e+01	-22.22247	< 2.2e-16	***
CityAurora	-27.8423	1.346e+01	-2.06822	0.03878	*
CityBoulder	36.0033	8.937e+01	0.40285	0.68711	
CityBrighton	83.6734	1.263e+02	0.66238	0.50782	
CityBroomfield	-15.7030	1.866e+01	-0.84151	0.40018	
CityCastle Pines	-77.2523	9.807e+01	-0.78773	0.43097	
CityCentennial	-15.5375	2.267e+01	-0.68552	0.49311	
CityCommerce City	-138.7205	4.951e+01	-2.80176	0.00514	**
CityDenver	45.8811	3.155e+01	1.45435	0.14604	
CityEdgewater	53.5573	5.441e+01	0.98441	0.32506	
CityEnglewood	22.0748	2.809e+01	0.78585	0.43207	
CityGolden	86.6280	6.442e+01	1.34469	0.17891	
CityGreenwood Village	-4.1622	4.492e+01	-0.09265	0.92619	
CityHenderson	-147.1701	1.618e+02	-0.90944	0.36325	
CityHighlands Ranch	-21.9284	3.877e+01	-0.56559	0.57175	
CityLafayette	-44.2774	6.911e+01	-0.64072	0.52179	
CityLakewood	32.2994	3.328e+01	0.97058	0.3319	
CityLittleton	-8.2044	2.686e+01	-0.30546	0.76005	
CityLone Tree	83.2722	1.378e+02	0.60425	0.54576	
CityLouisville	-39.6393	6.907e+01	-0.57386	0.56614	
CityMorrison	79.6842	8.769e+01	0.90870	0.36364	
CityNorthglenn	71.5007	4.995e+01	1.43131	0.15253	
CityParker	12.6482	3.575e+01	0.35384	0.7235	
CitySuperior	-62.4476	6.911e+01	-0.90359	0.36635	
CityThornton	77.4328	4.755e+01	1.62838	0.10364	
CityWestminster	-14.8513	2.141e+01	-0.69368	0.48798	
CityWheat Ridge	13.5252	2.735e+01	0.49454	0.62099	
ZIP	-0.2737	1.364e-01	-2.00729	0.04488	*
Store_Number	-1.9082	1.374e+00	-1.38881	0.16508	
Avg_Num_Products_Purchased	65.7662	1.885e+00	34.89620	< 2.2e-16	***

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

Fig 12: Analysis Report from Alteryx

Residual standard error: 136.29 on 1630 degrees of freedom
Multiple R-squared: 0.833, Adjusted R-Squared: 0.8297
F-statistic: 254.1 on 32 and 1630 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	19919762.99	3	357.45	< 2.2e-16 ***
City	544783.49	26	1.13	0.2985
ZIP	74846.11	1	4.03	0.04488 *
Store_Number	35828.73	1	1.93	0.16508
Avg_Num_Products_Purchased	22620637.95	1	1217.75	< 2.2e-16 ***
Residuals	30278618.62	1630		

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

Fig 13: ANOVA analysis report from Alteryx

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.

Using the ANOVA analysis report in fig 13, we see that the **customer_segment** and **Avg_num_products_purchased** variable show a lot of statistical significance (with **p-value < 2.2e^{-16}**) to the targe. ZIP shows little statistical significance (**p-value = 0.04488**) but it isn't enough to significally affect our model. The **R-squared value is 0.833** with an **adjusted R-Squared value of 0.8297**. This shows our model will be highly predictive for our target variable and our data is good enough to make predictions. I decided not to go back to the manager for more data. We can get better results by removing variables with little to no statistical significance.

Fig 14 below shows the results of removing all unsignificant variables, the **R-Squared value rose to 0.8353** and adjusted **R-Squared to 0.835.** This suggests that our model will be great for predicting this target variable.

: :Awesome: Good job using both R-squared and p-values to justify why your model is a good one! An r-squared of 0.835 means that about 84% of the target variable is explained by the predictor variables. In general, when a model with R-squared above 0.7 is considered a good model.

Variables that have a significant p-value (p

3. Validation

Report for Linear Model Avg_sale_amt_predictor

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-658.28	-67.13	-1.78	70.84	988.81

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	304.27	10.703	28.43	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.05	9.069	-16.43	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	272.18	12.043	22.60	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-246.12	9.872	-24.93	< 2.2e-16 ***
Avg_Num_Products_Purchased	66.58	1.534	43.40	< 2.2e-16 ***

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

Residual standard error: 136.93 on 2299 degrees of freedom
Multiple R-squared: 0.8353, Adjusted R-Squared: 0.835
F-statistic: 2915 on 4 and 2299 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	27050761.75	3	480.93	< 2.2e-16 ***
Avg_Num_Products_Purchased	35312939.48	1	1883.48	< 2.2e-16 ***
Residuals	43103338.82	2299		

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

Fig 14: Analysis report for final linear regression model

: :Awesome: You have all 3-star p-values, indicating strong significance.

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)

The analysis report above shows that the best equation to use for the model is:

Y = 304.27 – (149.05 * Customer_SegmentLoyalty Club Only) + (272.18 * Customer_SegmentLoyalty Club and Credit Card) – (246.12 * Customer_SegmentStore Mailing List) + (66.58 * Avg_Num_Products_Purchased)

Note: The base case is set to **Credit Card Only** for the customer segment variable.

: :Awesome: The linear equation is correct as it contains just the correct predictor variables with correct coefficients.

It missed only include zero coefficient for CreditCard segment. For the equation to be complete, it is important that all values of the CustomerSegment variable are represented in the equation.

Name	Customer_Segment	Avg_Num_Products_Purchased	Score_Yes	predicted_avg_sale_amt	Expected_Profit	Expected_revenue
A Giametti	Loyalty Club Only	3	0.305036	354.968291	47.639019	108.278038993363

Fig 15: Predicted results for A Giametti

Example:

Using customer A Gametti as an example,

Predicted_avg_sale_amount (A Gametti) = 304.27 – (149.05 x 1) + (272.18 x 0) – (246.12 x 0) + (66.58 x 3)

= 304.27 – 149.05 + 0 + 0 + 199.74

= 354.96

Thus, A Gametti's predicted_avg_sale_amt = 354.96

Step 3: Presentation/Visualization

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

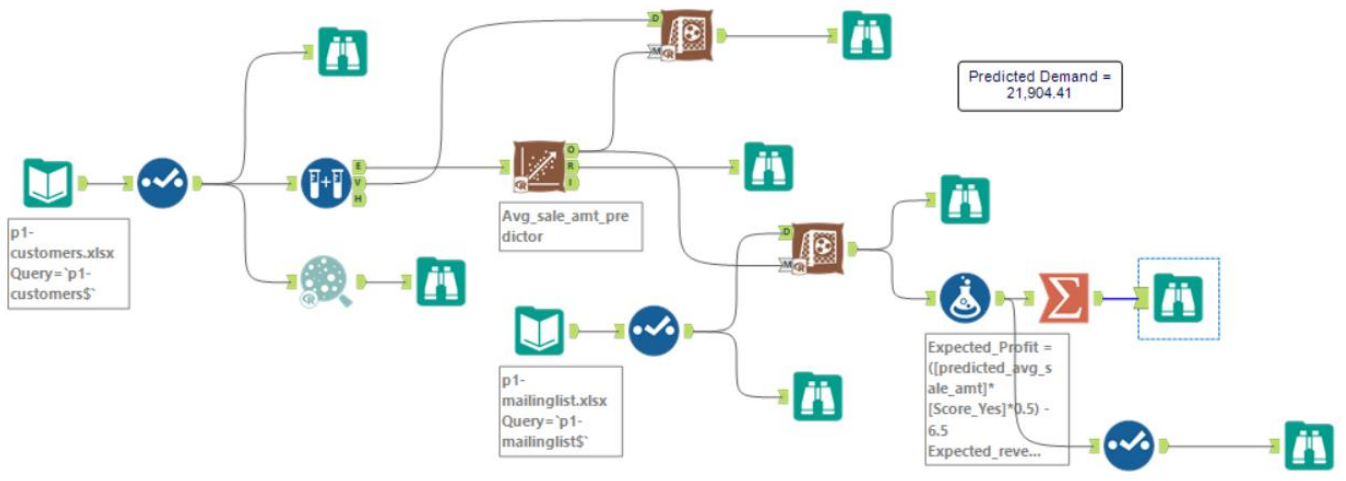

Fig 16: Alteryx Workflow used for this project

Sum_predicted_avg_sale_amt	Sum_Expected_revenue	Sum_Expected_Profit
137,806.101766	47,058.810715	21,904.405358

Fig 17: Sum of predicted_avg_sale_amt, expected_revenue and expected_profit

1. RECOMMENDATION

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

Management informed us that they would only send out the catalog if the expected profit was greater than $10,000. From my analysis, the profit would exceed the minimum expected profit, therefore, I recommend sending out the catalog to our new customers.

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

After creating the linear regression model, I used it to predict the average sale amount (predicted_avg_sale_amt) for the 250 customers in mailinglist.xlsx. Then I used the **formula tool** to calculate the expected_profit and expected_revenue of the customers using the formulas:

Expected_revenue = [predicted_avg_sale_amt] * [Score_Yes]

Expected_profit = ([predicted_revenue_per_customer] * 0.5) - 6.5

For example,

Using A Gametti who's predicted_avg_sale_amt was calculated to be 354.96 (shown in fig 15 above):

Expected_revenue (A Gametti) = [predicted_avg_sale_amt] * [Score_Yes]

= 354.96 x 0.305

= 108.27

Expected_profit (A Gametti) = ([predicted_revenue_per_customer] * 0.5) - 6.5

= (108.27 * 0.5) – 6.5

= 47.635

Thus, A Gametti's expected_revenue = 108.27 and expected_profit = 47.635

: :Awesome: Excellent job here thoroughly justifying how you arrived at the final recommendation. Exactly what you would want to present to management.

After calculating the expected revenue and profit, I used the **summarize tool** to calculate the sum of the predicted_avg_sale_amt, expected revenue, and expected profit. The catalog demand is the sum of expected profit which is **$21,904.4**.

2. Expected profit

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

In conclusion, the expected profit from the 250 customers is **$21,904.4**. This is significantly higher than the expected minimum profit of **$10,000**. This means that our company stands to make a lot of profit from sending this catalog and so I highly recommend the management send the catalog to the customers in the mailing list.

: : Awesome: The final expected profit is right!