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 decision my manager needs to make is to determine how much profit the company can expect from sending a catalog to these 250 new customers from their mailing list. The manager will only send the catalog out to these new customers if the expected profit contribution exceeds $10,000. As a business analyst, I am assigned to help my manager run the numbers because he is not very familiar with predictive models. I am tasked to predict the expected profit from these 250 new customers.

2. What data is needed to inform those decisions?

Statements:
i) The main task is to predict the profit from sending a catalog to the 250 new customers.
ii) The sales data from 250 new customers – from p1-mailinglist.xlsx - cannot be used to predict the possible profit from them. There is no actual sale record from them yet and the only field that is highly related to the predicted profit is the average number of products purchased.
iii) To predict the future profit from new customers, we need to gather the existing data from existing customers and these data we need must have certain impact on current profit from existing customers. We will use certain analysis and modeling on them to predict future possible profit from the new customers.
iii) The sales data from the existing customers are provided in p1-customers.xlsx. The field that is highly related to the profit from the existing customer is the average sale amount.
iv) In this case, we need to predict the average sale amount for each new customer.
v) From the average sale amount for each new customer, we multiply it with Score_Yes – the probability they would buy after they get a catalogue – and gross margin, and minus away $6.50 – the price for a catalogue – to calculate the predicted profit from each new 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.

At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) 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 this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

Below is the summary of what I would do.
From business context, I will select the suspected predictor variables that have influence on the target variable Avg_Sale_Amount and justify it with scatterplots and their linear relationships. With the chosen variables, I would use Alteryx to generate a Linear Regression model to predict the Avg_Sale_Amount for the next 250 customers.

Below are the unit tasks I would explain further to fulfill the summary above.

Below are the variables from past record, p1_customers.xlsx Excel file and some logic selection:

Predictor Variable	**Is Chosen? Reason.**	**Is Numerical ?**
Name	No. Different name does not affect target variable value.	-
Customer_Segment	Yes. Suspect to influence.	Text
Customer_ID	No. Customer ID is unique to each customer and does not affect target variable value.	-
Address	No. The addresses are dependent on city and they do not show any significant meaning in their text. Maybe a variable City representing an area would be chosen instead.	-
City	Yes. Suspect to influence.	Text
State	No. All the customers are from the same state, CO. They do not affect target variable value.	-
ZIP	Yes. Suspect to influence.	Numerical
Avg_Sale_Amount	No. This is the target variable we are going to predict.	-
Store_Number	Yes. Suspect to influence.	Numerical
Responded_to_last_Catalog	No. This will not be used because it cannot be applied to mailing list data set.	-
Avg_Num_Products_Purchased	Yes. Suspect to influence.	Numerical
#_Years_as_Customer	Yes. Suspect to influence.	Numerical

Table 2.1 *Variables from p1_customers.xlsx.*

From the chosen predictor variables that are numerical, I need to check if a numerical variable really impacts the target variable linearly, I used Alteryx to scatterplot to check for their linear relationships. Below is the result:

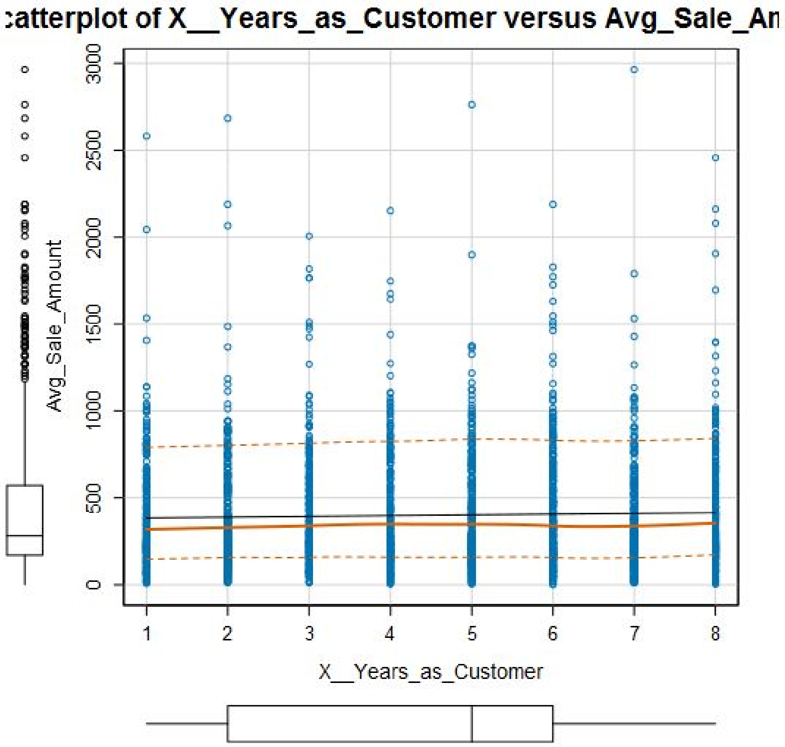
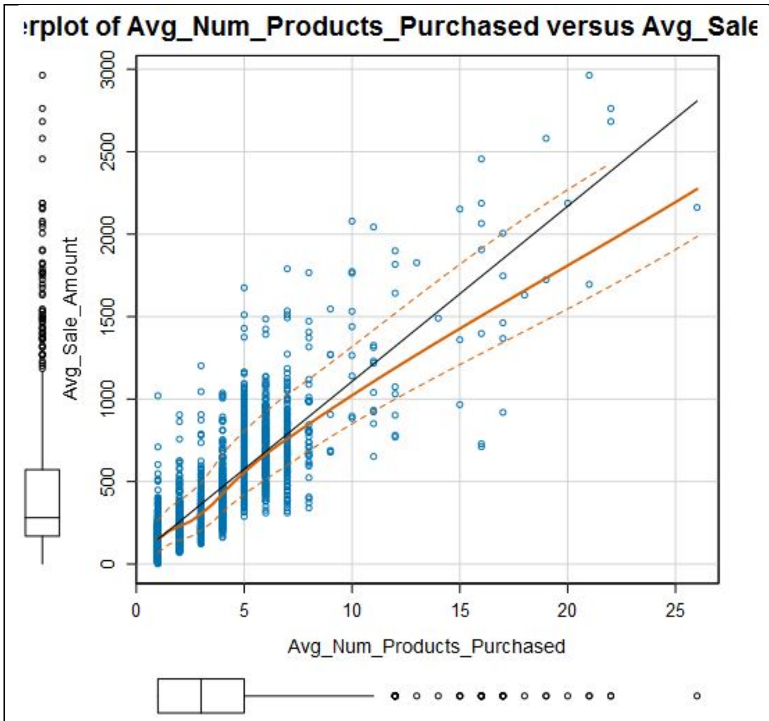
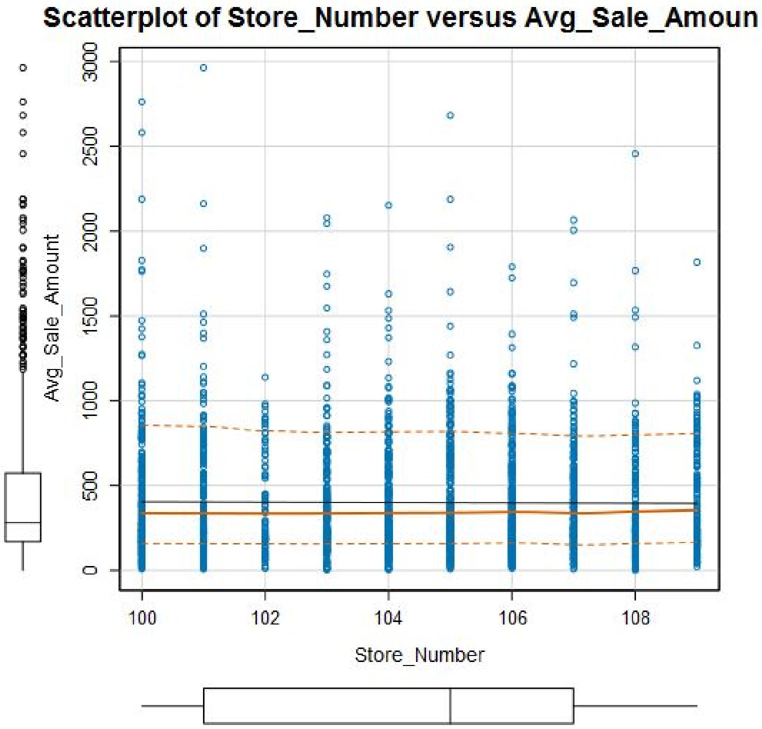


Table 2.2 *A scatterplot of Avg_Sale_Amount (y-axis) against #_Years_As_Customer (x-axis).*

This is a scatterplot of Avg_Sale_Amount (y-axis) against #_Years_As_Customer (x-axis).

 Table 2.3 *A scatterplot of Avg_Sale_Amount (y-axis) against Avg_Num_Products_Purchased (x-axis).*	This is a scatterplot of Avg_Sale_Amount (y-axis) against Avg_Num_Products_Purchased (x-axis).
 Table 2.4 *A scatterplot of Avg_Sale_Amount (y-axis) against Store_Number (x-axis).*	This is a scatterplot of Avg_Sale_Amount (y-axis) against Store_Number (x-axis).

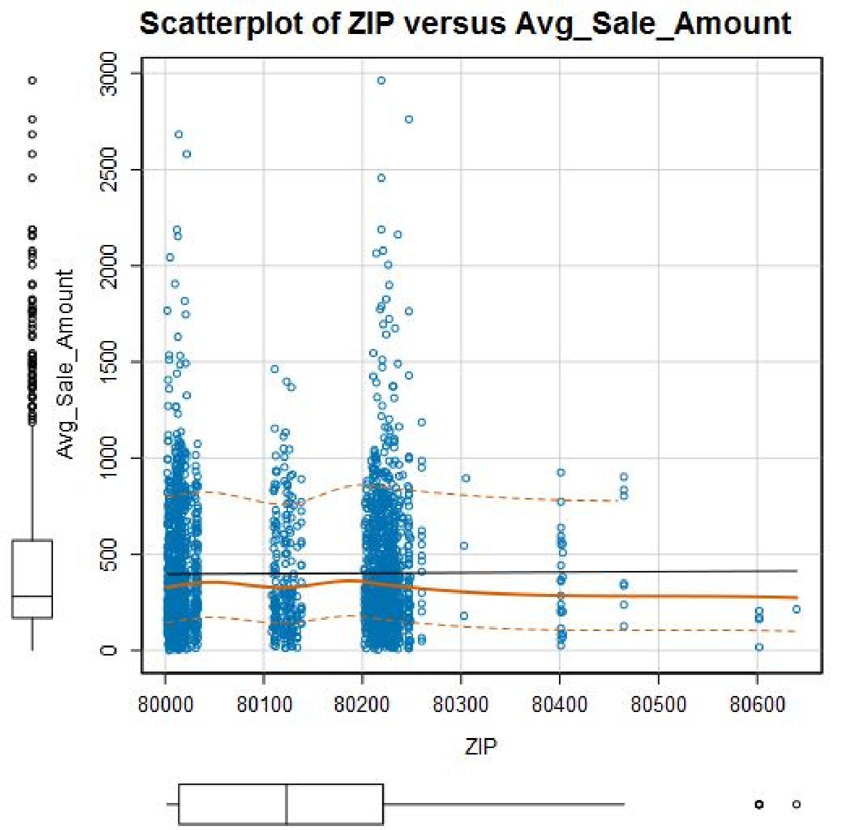


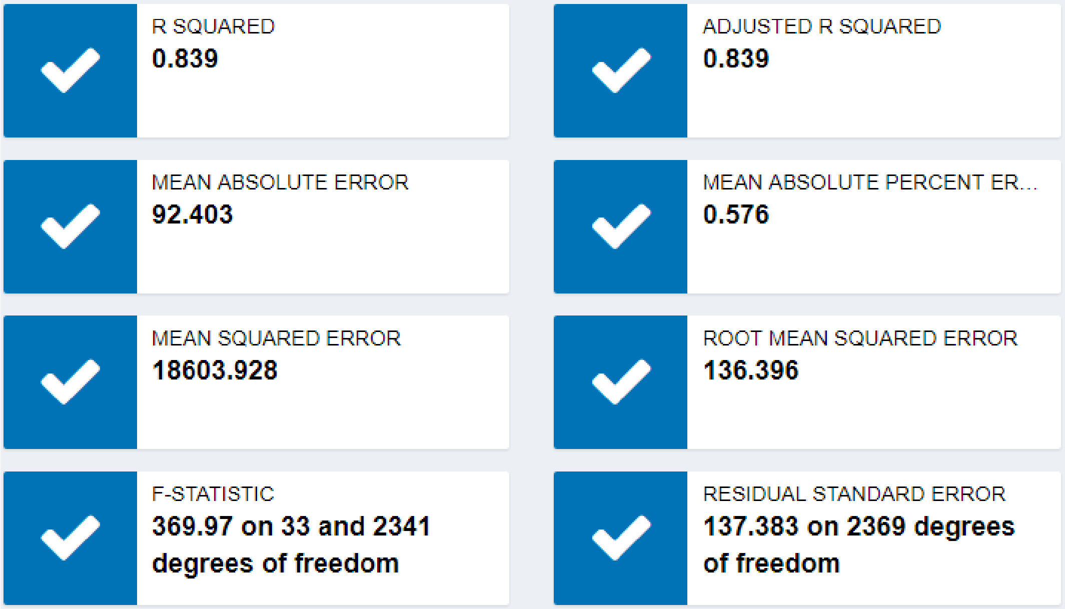
Table 2.5 *A scatterplot of Avg_Sale_Amount (y-axis) against ZIP (x-axis).*

This is a scatterplot of Avg_Sale_Amount (y-axis) against ZIP (x-axis).

From each of the scatterplots, only the plot of Avg_Sale_Amount (y-axis) against Avg_Num_Products_Purchased (x-axis) shows linear relationship. The rest of the scatterplots depict the relationship as categorical instead.

From all the chosen predictor variables (Table 2.1), I used Alteryx to generate a linear regression model.

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.



✓ R SQUARED
0.839

✓ ADJUSTED R SQUARED
0.839

✓ MEAN ABSOLUTE ERROR
92.403

✓ MEAN ABSOLUTE PERCENT ER...
0.576

✓ MEAN SQUARED ERROR
18603.928

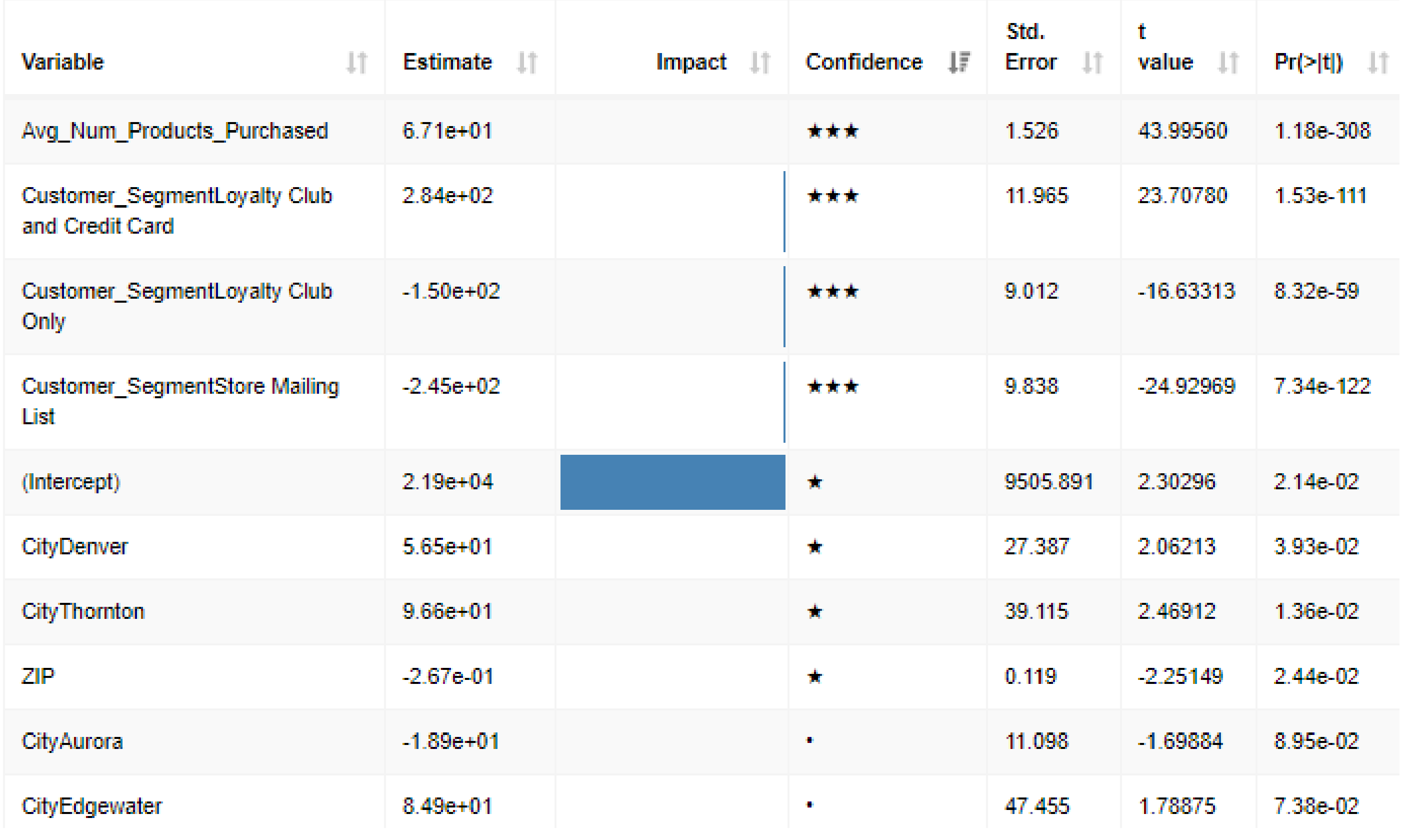
✓ ROOT MEAN SQUARED ERROR
136.396

✓ F-STATISTIC
369.97 on 33 and 2341 degrees of freedom

✓ RESIDUAL STANDARD ERROR
137.383 on 2369 degrees of freedom

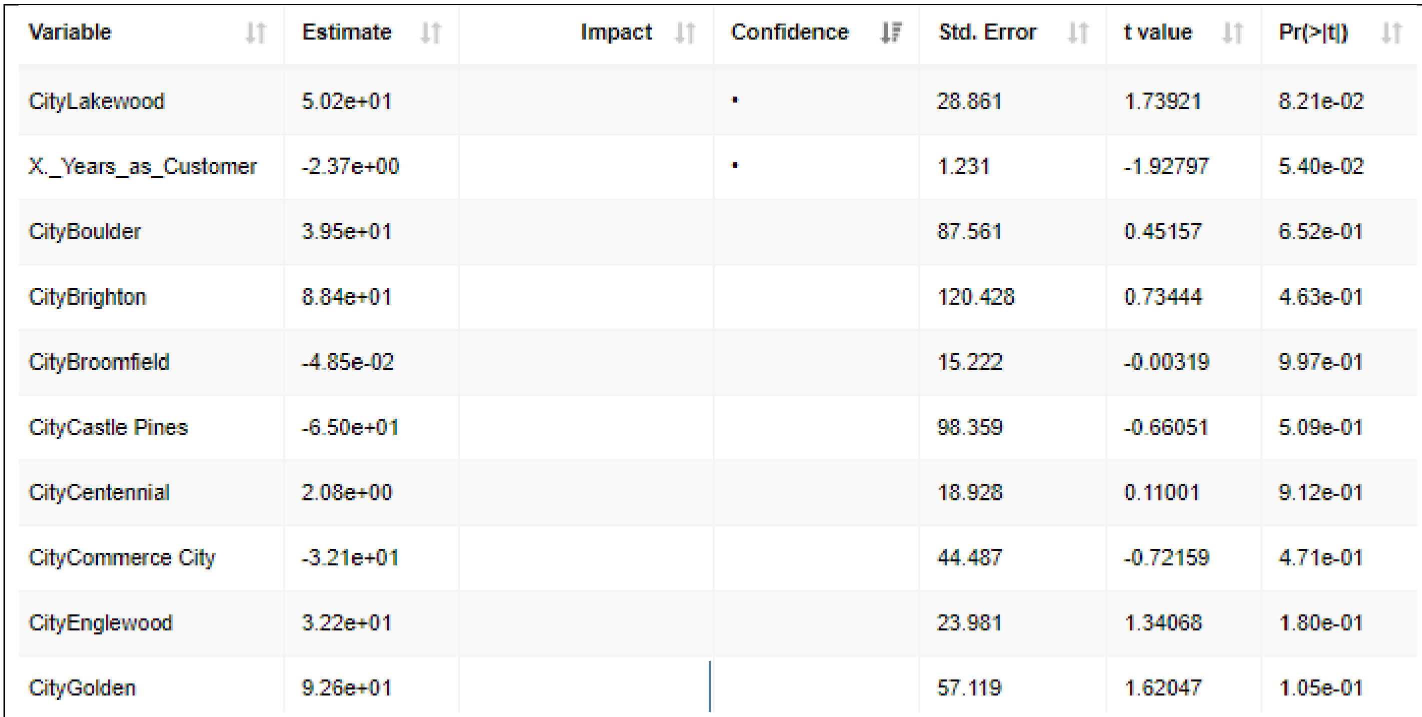
This is the generated report from the linear regression model that include all the predictor variables.

The high R-squared (the multiple and adjusted) value, with 0.839 (more than 0.7), indicates the model has nearly explained all variance in the target variable Avg_Sale_Amount. It has high explanatory power and thus is a good model.



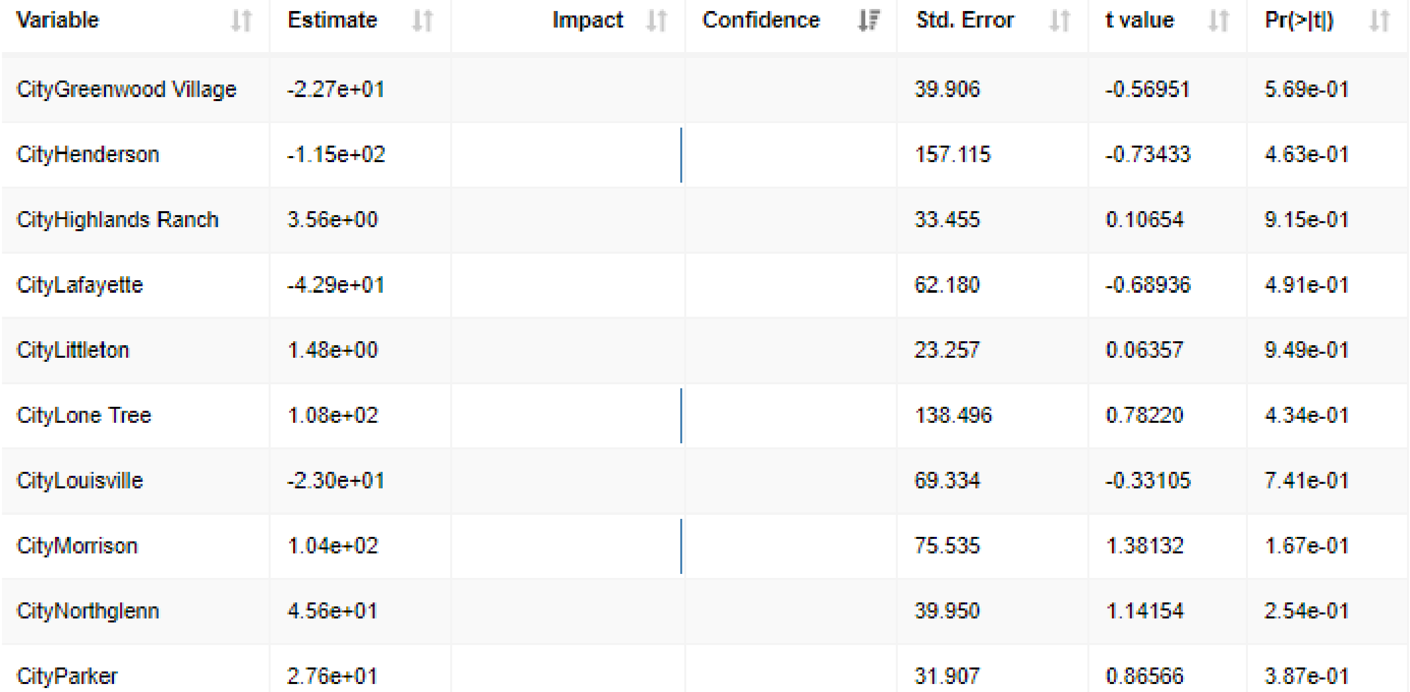
Variable	Estimate	Impact	Confidence	Std. Error	t value	Pr(>|t|)
Avg_Num_Products_Purchased	6.71e+01		★★★	1.526	43.99560	1.18e-308
Customer_SegmentLoyalty Club and Credit Card	2.84e+02		★★★	11.965	23.70780	1.53e-111
Customer_SegmentLoyalty Club Only	-1.50e+02		★★★	9.012	-16.63313	8.32e-59
Customer_SegmentStore Mailing List	-2.45e+02		★★★	9.838	-24.92969	7.34e-122
(Intercept)	2.19e+04		★	9505.891	2.30296	2.14e-02
CityDenver	5.65e+01		★	27.387	2.06213	3.93e-02
CityThornton	9.66e+01		★	39.115	2.46912	1.36e-02
ZIP	-2.67e-01		★	0.119	-2.25149	2.44e-02
CityAurora	-1.89e+01		•	11.098	-1.69884	8.95e-02
CityEdgewater	8.49e+01		•	47.455	1.78875	7.38e-02

This is the generated statistics page 1.



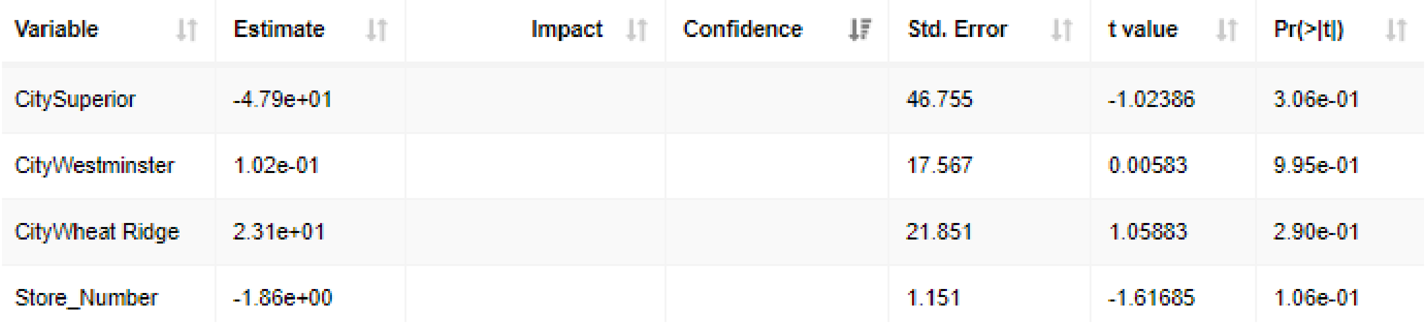
Variable	Estimate	Impact	Confidence	Std. Error	t value	Pr(>|t|)
CityLakewood	5.02e+01		.	28.861	1.73921	8.21e-02
X._Years_as_Customer	-2.37e+00		.	1.231	-1.92797	5.40e-02
CityBoulder	3.95e+01			87.561	0.45157	6.52e-01
CityBrighton	8.84e+01			120.428	0.73444	4.63e-01
CityBroomfield	-4.85e-02			15.222	-0.00319	9.97e-01
CityCastle Pines	-6.50e+01			98.359	-0.66051	5.09e-01
CityCentennial	2.08e+00			18.928	0.11001	9.12e-01
CityCommerce City	-3.21e+01			44.487	-0.72159	4.71e-01
CityEnglewood	3.22e+01			23.981	1.34068	1.80e-01
CityGolden	9.26e+01			57.119	1.62047	1.05e-01

This is the generated statistics page 2.



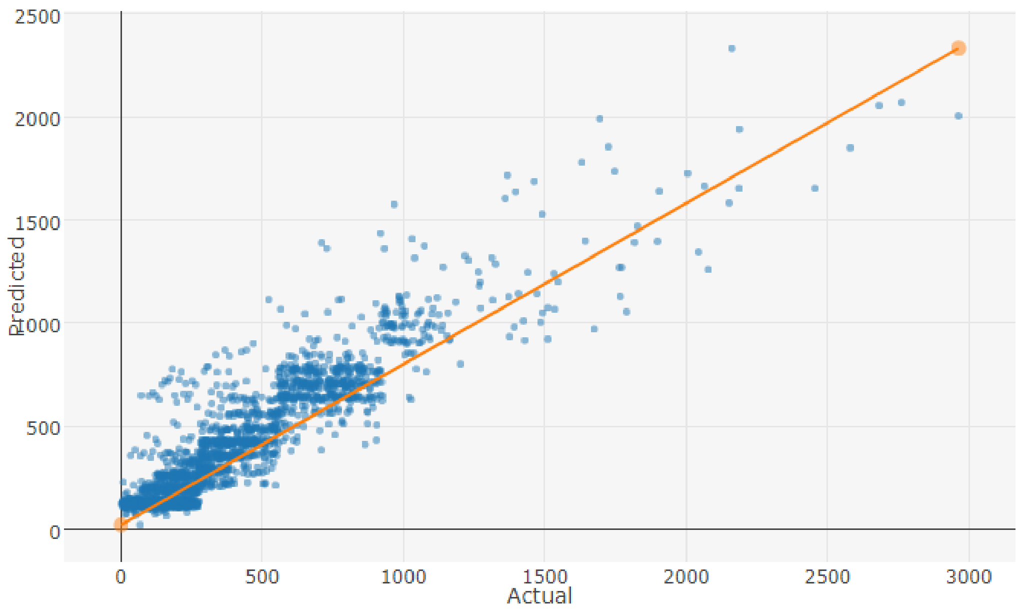
Variable	Estimate	Impact	Confidence	Std. Error	t value	Pr(>|t|)
CityGreenwood Village	-2.27e+01			39.906	-0.56951	5.69e-01
CityHenderson	-1.15e+02			157.115	-0.73433	4.63e-01
CityHighlands Ranch	3.56e+00			33.455	0.10654	9.15e-01
CityLafayette	-4.29e+01			62.180	-0.68936	4.91e-01
CityLittleton	1.48e+00			23.257	0.06357	9.49e-01
CityLone Tree	1.08e+02			138.496	0.78220	4.34e-01
CityLouisville	-2.30e+01			69.334	-0.33105	7.41e-01
CityMorrison	1.04e+02			75.535	1.38132	1.67e-01
CityNorthglenn	4.56e+01			39.950	1.14154	2.54e-01
CityParker	2.76e+01			31.907	0.86566	3.87e-01

This is the generated statistics page 3.

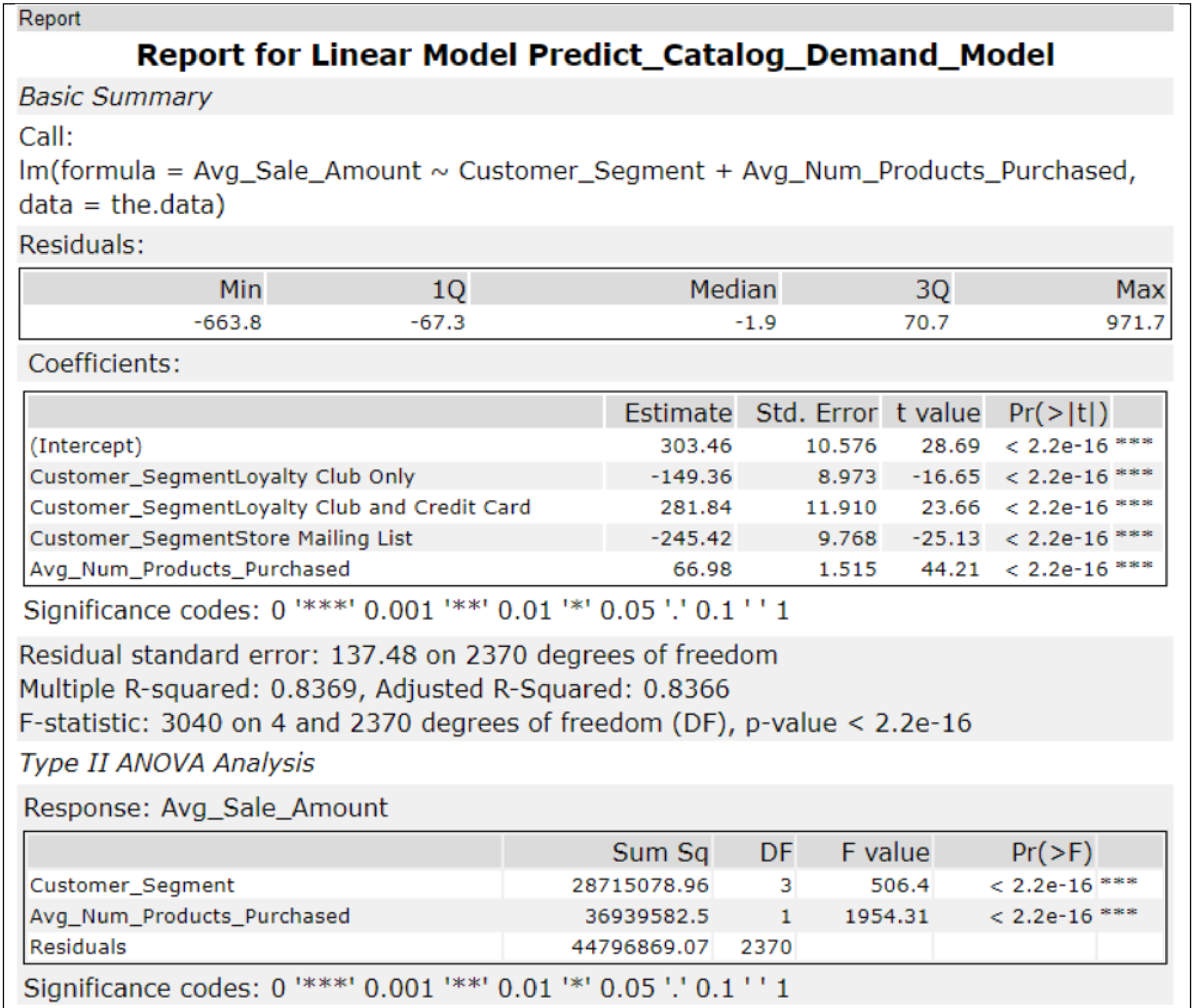


Variable	Estimate	Impact	Confidence	Std. Error	t value	Pr(>|t|)
CitySuperior	-4.79e+01			46.755	-1.02386	3.06e-01
CityWestminster	1.02e-01			17.567	0.00583	9.95e-01
CityWheat Ridge	2.31e+01			21.851	1.05883	2.90e-01
Store_Number	-1.86e+00			1.151	-1.61685	1.06e-01

This is the generated statistics page 4.

	This is a scatterplot of Predicted Avg_Sale_Amount (the orange best-fitted line) against Actual Avg_Sale_Amount (the blue dots).

The results above show the model that is generated from including all the predictor variables. To have a best model, I choose the predictor variables that have high Confidence level and most significant – Average Number of Products Purchased, Customer_Segment (Loyalty Club only), Customer Segment (Loyalty Club and Credit Card), Customer Segment (Store Mailing List). Next, I regenerate the linear regression model to get the best linear regression equation.



Report

Report for Linear Model Predict_Catalog_Demand_Model

Basic Summary

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

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

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

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Residuals	44796869.07	2370			

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

This is the generated report from the new linear regression model.

The high R-squared (the multiple and adjusted) value, with 0.837 (more than 0.7), indicates the model has nearly explained all variance in the target variable Avg_Sale_Amount. It has high explanatory power and thus is a good model.

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

From the generated statistic results obtained from Alteryx, the best linear regression equation I can get is here below:

Y = 303.46 + 66.98 * X1 – 149.36 * X2 + 281.84 * X3 – 245.42 * X4

where Y = Predicted Average Sale Amount,

X1 = Average Number of Products Purchased,

X2 = (If Customer Segment: Loyalty Club only),

X3 = (If Customer Segment: Loyalty Club and Credit Card),

X4 = (If Customer Segment: Store Mailing List),

Step 3: Presentation/Visualization

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

At the minimum, answer these questions:

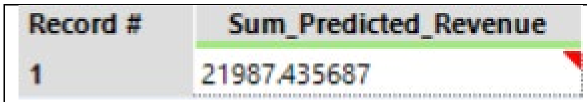
1. What is your recommendation? Should the company send the catalog to these 250 customers?

Yes.

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

i. I used Alteryx to produce the Linear Regression Model Statistics.
ii. From there, I chose the predictor variables that have low p-value (less than 0.05) to form a Linear Regression Equation.
iii. Using the equation, I used Alteryx to produce the predicted revenue for each customer.
iv. The predicted revenue formula is: [Predicted_Avg_Sale_Amount] * 0.5 * [Score_Yes] - 6.50
v. From the output, the predicted formula for all customer are positive.

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

 Record # Sum_Predicted_Revenue 1 21987.435687	This is the sum of the predicted revenue aka expected profit from the new catalog if sent to the 250 customers. The expected profit would be $21987.44.