Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   - Should the company send out this year's catalog to 250 new customers (based on the expected more than $10,000 profit)?

2. What data is needed to inform those decisions?
   - To make the decision, we need to calculate total expected profit contribution from 250 new customs. Before that we must predict the Avg_Sale_Amount for the same group.
   - We need historical data about the Sales from the last year when company send out its first catalog.
   - All the predictor variables which could have affected the Sales last year.
   - Formula to calculate the profit for the company.
   - Probability that the customer will buy the product after receiving the catalog.

Step 2: Analysis, Modeling, and Validation

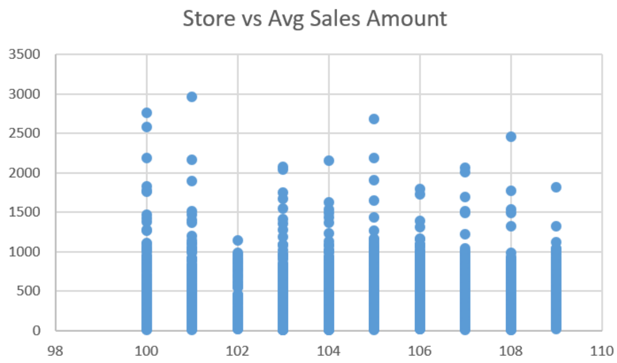
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable.

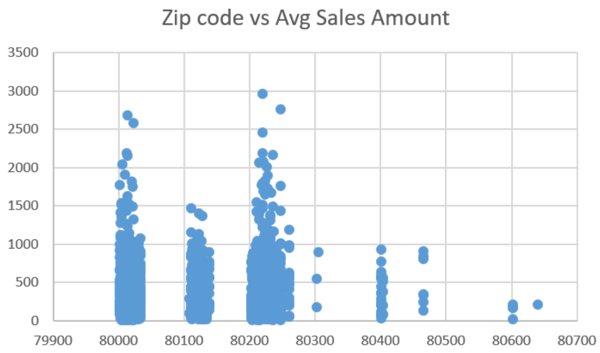
Answer: In order to predict the total profit from 250 new customers, it is important for us to understand from the existing customer data that which predictor variables have a linear relationship with the Avg_Sale_Amount. For the numerical variable we can simply create a scatter plot and understand the relationship and for categorical variable this can done by calculating p-value.

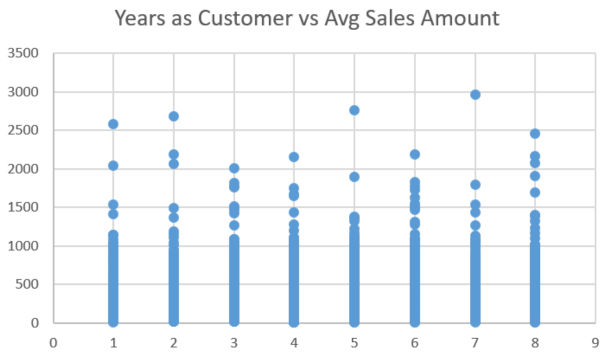
Useful Numerical predictor variables in the existing customer data:
Avg_Num_Products_Purchased, Years_as_Customer, Zip, Store Number

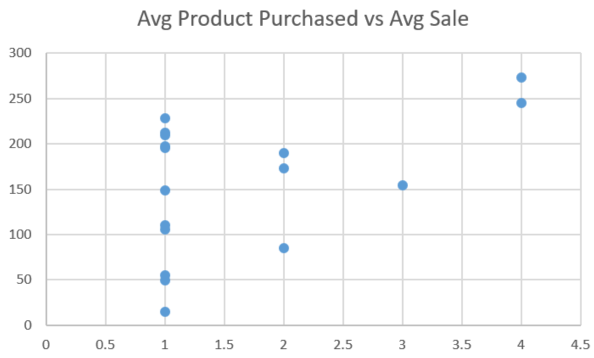
Useful Categorical predictor variables in the existing customer data:
Customer_Segment, City

Conclusion from the following scatterplots: Only Avg_Num_Products_Purchased has linear relationship with Avg_Sale_Amount. Going ahead with that, City and Customer_Segment.

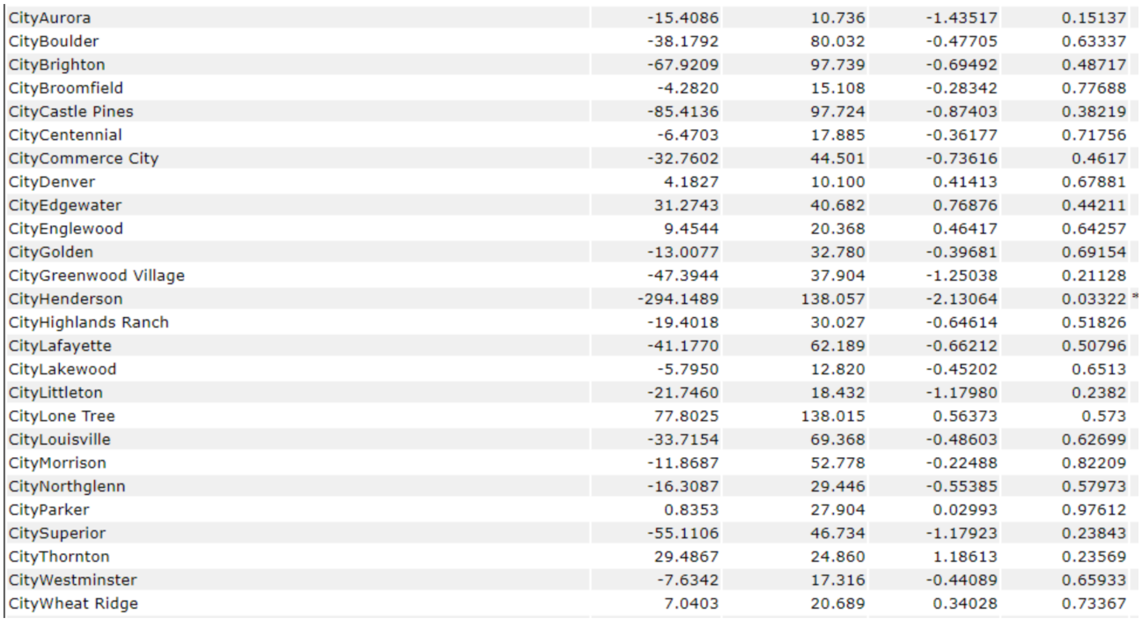

Store vs Avg Sales Amount


Zip code vs Avg Sales Amount


Years as Customer vs Avg Sales Amount


Avg Product Purchased vs Avg Sale

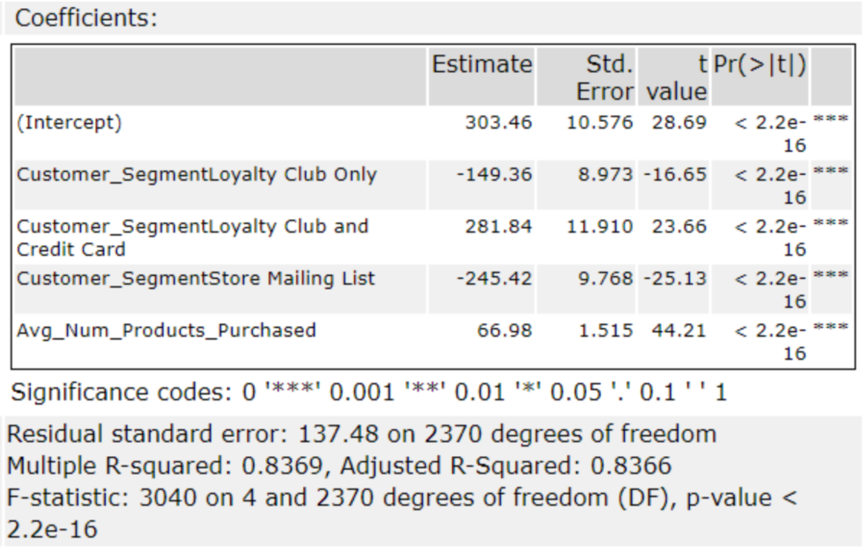
On the basis of the p-value for the City only Avg_Num_Products_Purchased and Customer_Segment are significant for our model.



CityAurora	-15.4086	10.736	-1.43517	0.15137
CityBoulder	-38.1792	80.032	-0.47705	0.63337
CityBrighton	-67.9209	97.739	-0.69492	0.48717
CityBroomfield	-4.2820	15.108	-0.28342	0.77688
CityCastle Pines	-85.4136	97.724	-0.87403	0.38219
CityCentennial	-6.4703	17.885	-0.36177	0.71756
CityCommerce City	-32.7602	44.501	-0.73616	0.4617
CityDenver	4.1827	10.100	0.41413	0.67881
CityEdgewater	31.2743	40.682	0.76876	0.44211
CityEnglewood	9.4544	20.368	0.46417	0.64257
CityGolden	-13.0077	32.780	-0.39681	0.69154
CityGreenwood Village	-47.3944	37.904	-1.25038	0.21128
CityHenderson	-294.1489	138.057	-2.13064	0.03322 *
CityHighlands Ranch	-19.4018	30.027	-0.64614	0.51826
CityLafayette	-41.1770	62.189	-0.66212	0.50796
CityLakewood	-5.7950	12.820	-0.45202	0.6513
CityLittleton	-21.7460	18.432	-1.17980	0.2382
CityLone Tree	77.8025	138.015	0.56373	0.573
CityLouisville	-33.7154	69.368	-0.48603	0.62699
CityMorrison	-11.8687	52.778	-0.22488	0.82209
CityNorthglenn	-16.3087	29.446	-0.55385	0.57973
CityParker	0.8353	27.904	0.02993	0.97612
CitySuperior	-55.1106	46.734	-1.17923	0.23843
CityThornton	29.4867	24.860	1.18613	0.23569
CityWestminster	-7.6342	17.316	-0.44089	0.65933
CityWheat Ridge	7.0403	20.689	0.34028	0.73367

2. Explain why you believe your linear model is a good model.

Answer: Predictors variable considered for analysis are: Avg_Num_Products_Purchased and Customer_Segment. Following is the results from the Linear Regression model.



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

Here is the reason, why this model is a good model:

- P-value for all the predictors variables are way less than 0.05, which means probability that coefficient for Customer_Segment and Avg_Num_Products_Purchased are zero is very less. That indicates that the model is significant.

- R-squared value for the model is .8369 which is close to 1. It means data fits very well in the created 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)

Answer:

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) + 0 * Customer_Segment (Credit_Card_Only) + 66.98 * Avg_Number_Products_Purchased

Step 3: Presentation/Visualization

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

Answer: The company should send the catalog to 250 new customers because the predicted overall profit is more than $10,000.

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

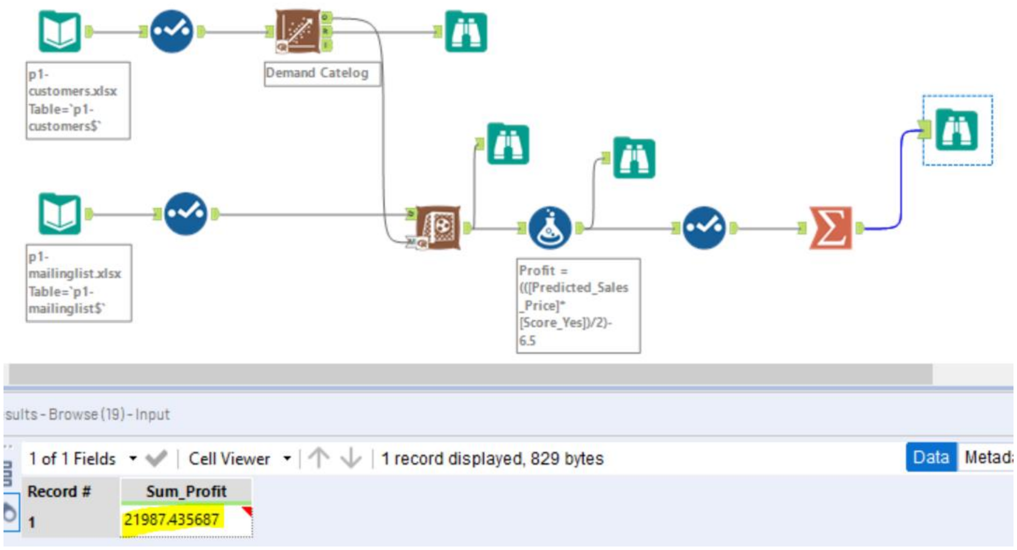
Step 1: Predicted the Avg_Sales_Amount for 250 new customers using the linear regression model.

Step 2: Calculated the Expected profit as per the details provided in the project description:
Expected_Profit = (([Avg_Sales_Amount] * [Score_Yes]) * .50) – [6.5]

Step 3: Expected profit is greater than $10,000. Hence, its good to send the catalog.

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

Answer: $ 21,987.435



1 record displayed, 829 bytes

Record #	Sum_Profit
1	21987.435687