Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

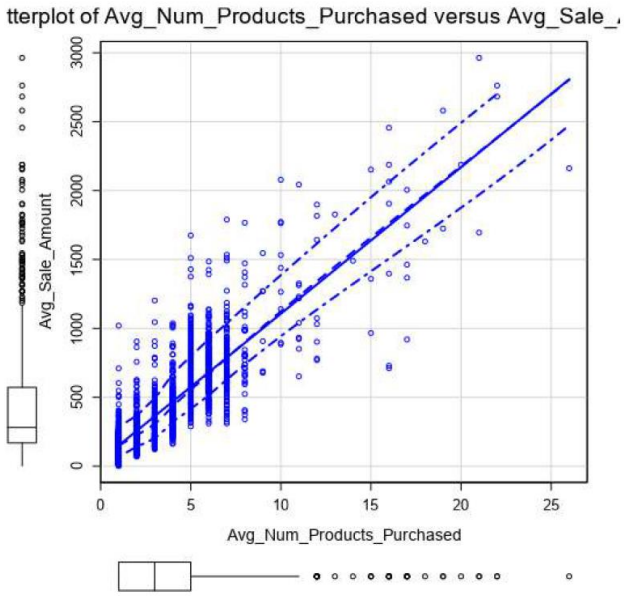
1. **What decisions needs to be made?**
   - Should the Company send out this year's catalog to 250 new customers (Based on the expected more than $10,000 Profit).

2. **What data is needed to inform those decisions?**
   - We are given two files of dataset i.e. customers.xlsx and mailing.xlsx. We need Avg_Num_Products_Purchased, Customer_Segment, Score_Yes, Mailing, Cost of Catalogue (i.e. $6.50) and Gross margin (i.e. 50%) to find the Profit.

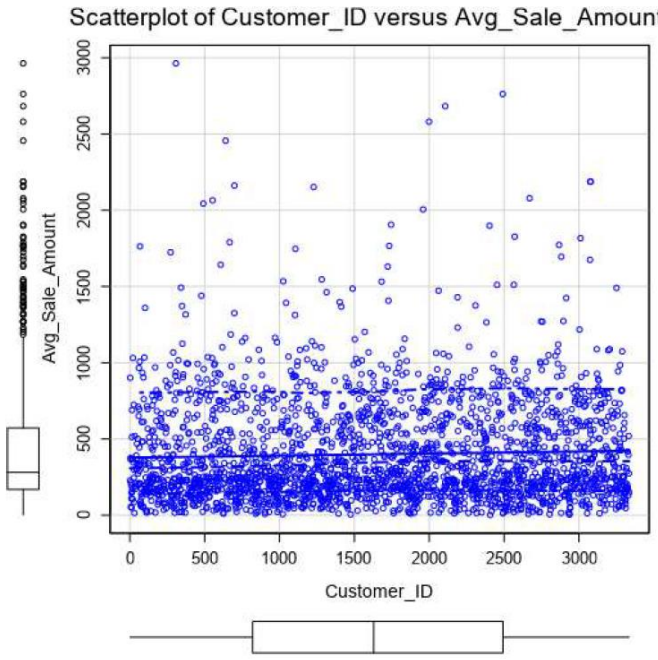
Step 2: Analysis, Modeling, and Validation

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. Please refer back 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.**

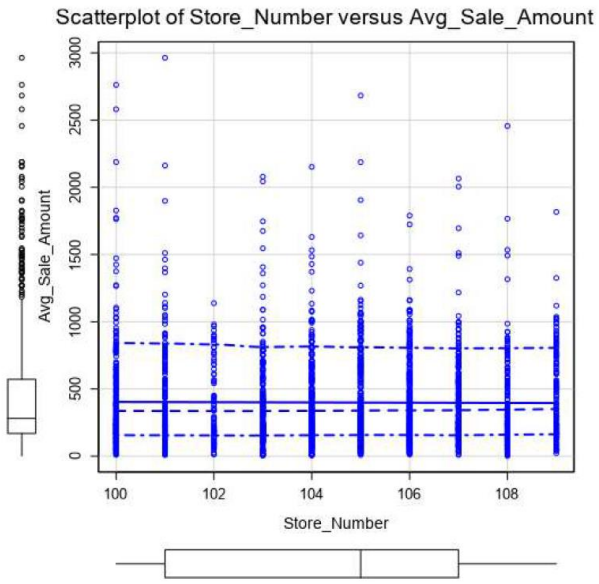
   - In order to Predict the Total Profit from 250 new customers, it is important for us to understand the data (i.e. Customer) which is given to us and from that which predictor variables have linear relationship with Avg_Sale_Amount.
   - For Numerical Variable we can simply create a Scatter Plot to understand the Relationship between them. For Categorical Variable we can simply use p-value.
   - In the Given Data useful Numerical Variables are: Avg_Num_Products_Purchased, Year_as_Customer, Zip, Store_Number, Customer_ID.
   - And for the Categorical Variable are: Customer_Segment and City.



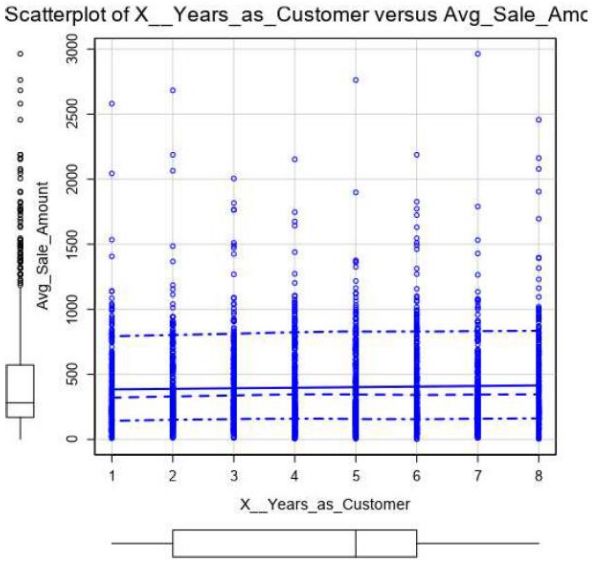
Scatter Plot of Avg_Num_Products_Purchased Vs Avg_Sale_Amount



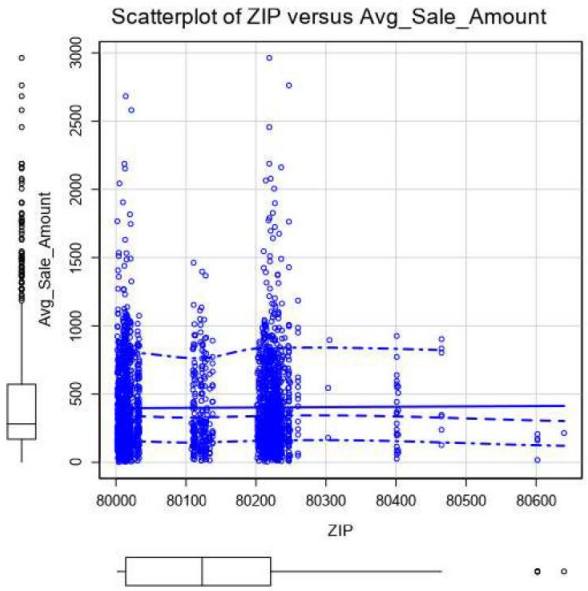
Scatter Plot of Customer_ID Vs Avg_Sale_Amount



Scatter Plot of Store_Number Vs Avg_Sale_Amount



Scatter Plot of Year_as_Customer Vs Avg_Sale_Amount



Scatter Plot of Zip Vs Avg_Sale_Amount

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	307.1425	13.448	22.83890	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.7086	9.020	-16.59807	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	281.9319	11.972	23.54995	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-244.9898	9.848	-24.87665	< 2.2e-16 ***
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
Avg_Num_Products_Purchased	67.1321	1.527	43.95115	< 2.2e-16 ***

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

- On the Basis of Scatter Plot, we can say that only Avg_Num_Products_Purchased has linear relationship with Avg_Sale_Amount. From P-value we can state that Customer_Segment and Avg_Num_Products_Purchased are only significant for our 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.**
    - ❖ Predictors Variable are considered for analysis are: Avg_Num_Products_Purchased and Customer_Segment. Results from our Linear Regression model are:

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

    - ❖ From above Table, we can say that our model is good by considering these points:
        - o P-Value of all the predictors variable are way less than 0.05, which means probability of coefficient of Avg_Num_Products_Purchased and Customer_Segment are very less this indicates that model is significant.
        - o R-squared value for the model is 0.8369 which is very close to 1 it indicates that 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)**

    - ❖ 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: Mailing List) + 66.98 (Avg_Num_Products_Purchased).

Step 3: Presentation/Visualization

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

    - ❖ Yes, the Company Should send the Catalog to 250 new customers.

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

   - I have come to this recommendation based on these steps:
     - **Step 1:** Predicted the Avg_Sales_Amount for **250** new customers using the linear regression model.
     - **Step 2:** Calculated the expected profit as per project description:
       Profit = (([Predicted_Sales_Price] * [Score_Yes]) / 2) - 6.5
     - **Step 3:** Profit is Greater than $10,000. Hence, it is good to send the catalog.

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

   - **The Expected Profit from the New Catalog is: $21987.4357**