Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Project 1: Predicting Catalog Demand

The Business Problem

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Step 1: Business and Data Understanding

1. What decisions needs to be made?

The decision that needs to be made throughout this analysis is whether the predicted profit from 250 new customers would exceed $10,000 for the new catalog launch, based on the previous catalog sales data.

2. What data is needed to inform those decisions?

To predict the profit of the new catalog launch, we need previous sales data of the existing customers, and the variables that have somewhat relationship to average sales amount. Since we do have the previous average sales data of the customers, we are data rich. We are projecting the average sales price of the new catalog launch, so we use linear regression model because the average sales amount that we want to predict is numeric and continuous.

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.

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

First, a linear model was created on all possible variables against the target variable, average sale amount. As it can be seen in the below table, only the variables "Customer Segment", "Store_Number", and "Average_Num_ Products_Purchased" showed statistical significance with p-value less than 0.05. Variable "No_Years_as_customer" had p-value close to 0.05 and it seemed to have some power in explaining the target variable, but as we are guided to only select variables with high significance level (p-value less than 0.05), this variable was not selected.

Then I created scatterplots on all the significant explanatory variable vs the target variable to detect any linear-like relationships in the variables. Since only numeric variables could be scatter-plotted in Alteryx, the scatterplot of "Customer_Segment" vs "Avg_Sale_Amount" was not performed.

Record	Report
1	**Report for Linear Model catalog_demand_history_lm**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Customer_ID + City + ZIP + Store_Number + Avg_Num_Products_Purchased + No_Years_as_customer, data = the.data)
	Residuals:

Min	1Q	Median	3Q	Max
-6.87e+02	-6.73e+01	-1.02e-13	6.84e+01	9.43e+02

Coefficients: (8 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.165e+02	140.181	4.398234	1e-05	***
Customer_SegmentLoyalty Club Only	-1.488e+02	9.176	-16.216437	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.874e+02	12.277	23.407996	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.442e+02	10.008	-24.405071	< 2.2e-16	***
Customer_ID	-1.764e-03	0.003	-0.587813	0.55672	
CityAurora	-2.174e+01	212.688	-0.102200	0.91861	

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Record Report

CityBoulder	-9.913e+01	140.367	-0.706213	0.48013	
CityBrighton	1.558e+01	217.005	0.071809	0.94276	
CityBroomfield	-2.759e+01	195.316	-0.141279	0.88766	
CityCastle Pines	-1.165e+02	100.793	-1.156153	0.24774	
CityCentennial	-3.038e+00	214.391	-0.014171	0.9887	
CityCommerce City	-5.760e+01	50.304	-1.144955	0.25235	
CityDenver	5.746e+00	161.409	0.035598	0.97161	
CityEdgewater	4.075e+01	164.790	0.247302	0.8047	
CityEnglewood	2.823e+00	221.854	0.012723	0.98985	
CityGolden	2.759e+01	119.380	0.231112	0.81725	
CityGreenwood Village	-1.011e+02	224.169	-0.451084	0.65197	
CityHenderson	-3.022e+02	140.301	-2.154029	0.03134	*
CityHighlands Ranch	2.987e+01	297.414	0.100449	0.92	
CityLafayette	-6.709e+01	66.591	-1.007434	0.31383	
CityLakewood	3.080e+01	158.283	0.194565	0.84575	
CityLittleton	5.929e+01	259.001	0.228901	0.81897	
CityLone Tree	1.373e+02	324.382	0.423116	0.67225	
CityLouisville	-4.546e+01	73.295	-0.620215	0.53518	
CityMorrison	-4.068e+01	57.920	-0.702294	0.48257	
CityNorthglenn	-4.208e+01	168.343	-0.249976	0.80263	
CityParker	4.100e+00	43.741	0.093724	0.92534	
CitySuperior	-7.124e+01	52.224	-1.364084	0.17268	
CityThornton	6.502e+01	167.466	0.388281	0.69784	
CityWestminster	-2.129e+01	192.225	-0.110743	0.91183	
CityWheat Ridge	-8.710e+01	166.025	-0.524617	0.5999	
ZIP80003	-3.155e+01	30.969	-1.018925	0.30835	
ZIP80004	-2.347e+01	29.178	-0.804464	0.42121	

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Record Report

ZIP80005	-1.578e+01	29.755	-0.530318	0.59594	
ZIP80007	-1.651e+01	57.682	-0.286213	0.77474	
ZIP80010	1.674e+00	215.750	0.007757	0.99381	
ZIP80011	-6.604e+00	215.158	-0.030692	0.97552	
ZIP80012	-3.171e+01	213.629	-0.148453	0.882	
ZIP80013	-2.880e+01	214.474	-0.134298	0.89318	
ZIP80014	3.567e+00	214.765	0.016607	0.98675	
ZIP80015	-1.025e+01	215.010	-0.047664	0.96199	
ZIP80016	-3.542e+01	215.272	-0.164541	0.86932	
ZIP80017	-2.901e+01	214.947	-0.134957	0.89266	
ZIP80018	-1.708e+02	228.540	-0.747277	0.45497	
ZIP80020	1.968e+01	196.301	0.100260	0.92015	
ZIP80021	-1.939e+01	197.879	-0.097988	0.92195	
ZIP80023	2.133e+01	200.543	0.106349	0.91531	
ZIP80030	2.873e+01	198.206	0.144938	0.88477	
ZIP80031	-2.306e+01	194.737	-0.118406	0.90576	
ZIP80033	9.739e+01	169.115	0.575906	0.56474	
ZIP80110	-2.081e+01	225.504	-0.092292	0.92647	
ZIP80111	3.512e+01	224.332	0.156544	0.87562	
ZIP80112	-6.605e+01	219.104	-0.301446	0.7631	
ZIP80113	-2.864e+01	225.159	-0.127178	0.89881	
ZIP80120	-1.289e+02	262.581	-0.490913	0.62354	
ZIP80121	-1.547e+01	220.577	-0.070133	0.94409	
ZIP80122	-9.071e+01	220.649	-0.411095	0.68104	
ZIP80123	-9.189e+01	261.501	-0.351412	0.72531	

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Record Report

ZIP80124	-7.990e+01	294.713	-0.271104	0.78634	
ZIP80126	-1.304e+01	294.431	-0.044297	0.96467	
ZIP80127	-1.152e+02	262.782	-0.438312	0.6612	
ZIP80128	-1.461e+02	263.203	-0.554966	0.57897	
ZIP80129	-1.017e+02	302.414	-0.336451	0.73656	
ZIP80130	-1.366e+02	302.954	-0.451014	0.65202	
ZIP80134	-7.534e+01	53.422	-1.410250	0.1586	
ZIP80202	-3.345e+01	168.498	-0.198535	0.84264	
ZIP80203	-2.439e+01	168.125	-0.145098	0.88465	
ZIP80204	-5.002e+01	165.555	-0.302112	0.76259	
ZIP80205	-4.425e-01	165.836	-0.002668	0.99787	
ZIP80206	-3.959e+01	167.423	-0.236492	0.81307	
ZIP80207	-4.475e+01	166.934	-0.268077	0.78866	
ZIP80209	-6.361e+01	166.770	-0.381412	0.70293	
ZIP80210	-6.273e+01	165.483	-0.379092	0.70466	
ZIP80211	5.580e+00	165.066	0.033806	0.97303	
ZIP80212	-1.900e+01	164.233	-0.115686	0.90791	
ZIP80214	-3.073e+01	161.802	-0.189942	0.84937	
ZIP80215	-9.948e+01	162.558	-0.611952	0.54063	
ZIP80216	2.017e+01	169.595	0.118915	0.90535	
ZIP80218	6.397e+00	166.956	0.038318	0.96944	
ZIP80219	-1.955e+01	163.984	-0.119219	0.90511	
ZIP80220	-7.894e+00	164.825	-0.047890	0.96181	

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Record Report

ZIP80221	-1.405e+01	165.352	-0.084954	0.93231	
ZIP80222	-3.833e+01	165.226	-0.232010	0.81655	
ZIP80223	-4.789e+01	168.113	-0.284895	0.77575	
ZIP80224	-1.593e+00	165.276	-0.009637	0.99231	
ZIP80226	-6.028e+01	161.030	-0.374317	0.7082	
ZIP80227	-2.733e+01	162.323	-0.168362	0.86631	
ZIP80228	-8.381e+01	162.647	-0.515308	0.60639	
ZIP80229	-5.059e+01	167.972	-0.301169	0.76331	
ZIP80230	-1.935e+01	166.779	-0.116031	0.90764	
ZIP80231	-3.417e+01	164.729	-0.207403	0.83571	
ZIP80232	-6.657e+01	161.782	-0.411472	0.68077	
ZIP80233	9.860e+00	171.858	0.057372	0.95425	
ZIP80234	1.529e+01	167.587	0.091235	0.92731	
ZIP80235	-3.945e+01	166.828	-0.236464	0.81309	
ZIP80236	-8.330e+01	165.205	-0.504246	0.61414	
ZIP80237	-4.792e+01	165.837	-0.288952	0.77264	
ZIP80238	-2.297e+01	170.450	-0.134774	0.8928	
ZIP80239	-1.066e+02	171.546	-0.621294	0.53447	
ZIP80241	-4.076e+01	173.867	-0.234438	0.81467	
ZIP80246	-6.145e+01	167.300	-0.367281	0.71344	
ZIP80247	-2.104e+01	164.409	-0.127967	0.89819	
ZIP80249	-2.958e+01	174.550	-0.169489	0.86543	
ZIP80260	-1.079e+02	169.555	-0.636252	0.52468	

Record Report

ZIP80303	6.276e+01	169.030	0.371266	0.71047	
ZIP80401	-7.111e+01	127.037	-0.559744	0.57571	
ZIP80403	-1.682e+01	100.566	-0.167232	0.8672	
ZIP80602	-1.056e+02	195.495	-0.540204	0.58911	
Store_Number	-2.608e+00	1.283	-2.033059	0.04216	*
Avg_Num_Products_Purchased	6.721e+01	1.557	43.162064	< 2.2e-16	***
No_Years_as_customer	-2.428e+00	1.261	-1.925203	0.05433	.
ZIP80022	NA	NA	NA	NA	
ZIP80026	NA	NA	NA	NA	
ZIP80027	NA	NA	NA	NA	
ZIP80108	NA	NA	NA	NA	
ZIP80138	NA	NA	NA	NA	
ZIP80305	NA	NA	NA	NA	
ZIP80465	NA	NA	NA	NA	
ZIP80640	NA	NA	NA	NA	

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

8 Residual standard error: 137.8 on 2264 degrees of freedom
Multiple R-squared: 0.8434, Adjusted R-Squared: 0.8358
F-statistic: 110.9 on 110 and 2264 degrees of freedom (DF), p-value < 2.2e-16

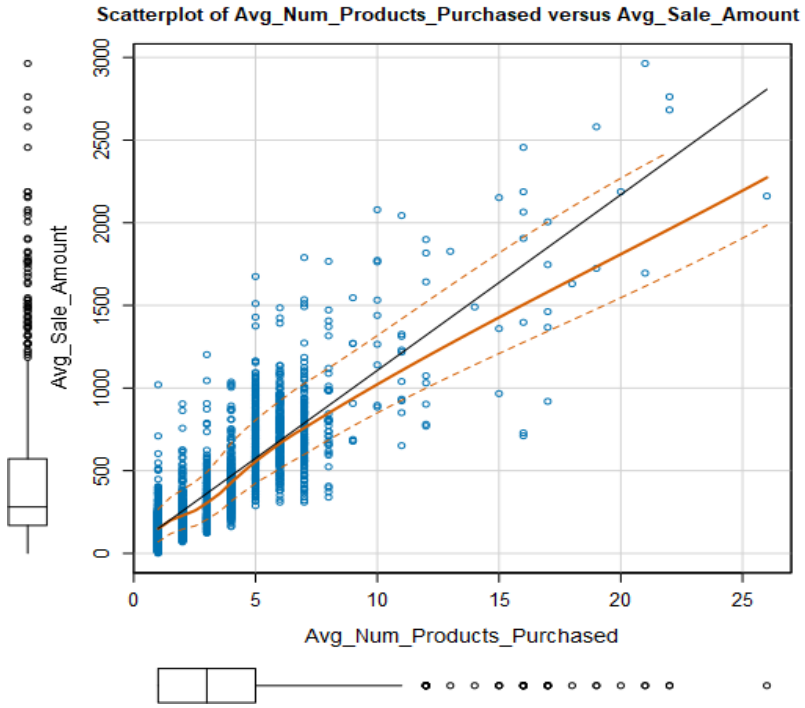
9 *Type II ANOVA Analysis*

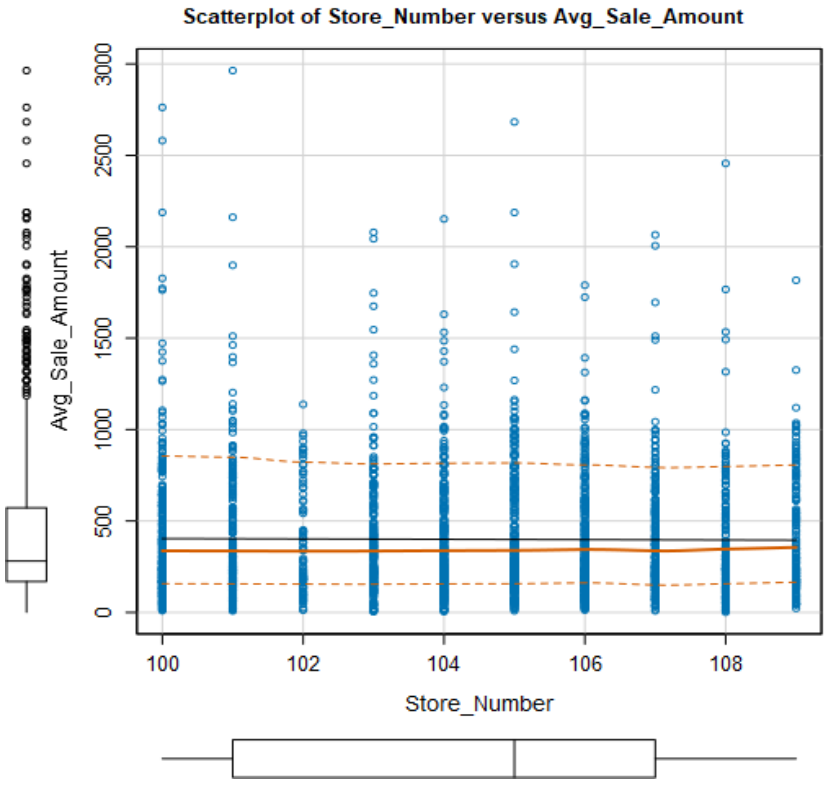
10 Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	27584851.37	3	484.2	< 2.2e-16	***
Customer_ID	6561.54	1	0.35	0.55672	
City	452332.57	26	0.92	0.58624	
ZIP	1282024.93	77	0.88	0.76873	
Store_Number	78492.31	1	4.13	0.04216	*
Avg_Num_Products_Purchased	35377853.34	1	1862.96	< 2.2e-16	***
No_Years_as_customer	70385	1	3.71	0.05433	.
Residuals	42993568.6	2264			

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

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree





As it can clearly be seen in the above scatterplots, only "Avg_Num_Products_Purchased" variable shows linear relationship to the target variable, but "Store_Number" variable does not. Therefore, our final model should only contain two explanatory variables: "Customer_Segment", and "Avg_Num_Products_Purchased"

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.

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

In general, the final regression model seems to well illustrate the linear relationship between predictor variables and average sales amount. Both Multiple R-Squared and Adjusted R-Squared value exceeds 0.8, which shows decent explanatory power of the model, and each explanatory variable is statistically significant with P-value all less than 0.05.

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)

The final regression equation is:
Y = 303.46 – 149.36*(If Customer Segment: Loyal Type Only) + 281.84*(If Customer Segment: Loyalty Club and Credit Card) – 245.42*(If Customer Segment: Store Mailing List) + 66.98*(Average Number of Products Purchased)

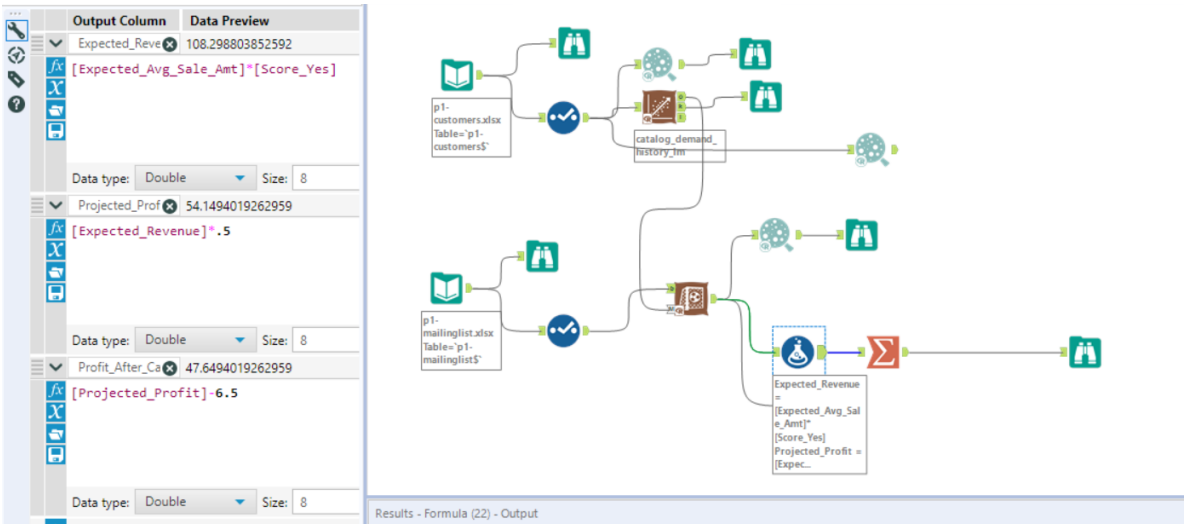
Step 3: Presentation/Visualization

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

My final recommendation is to proceed the production and send the catalog to these new 250 customers since the expected profit exceeds $10,000.

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

First, the expected revenue from sending the catalogs to 250 customers was calculated by multiplying average sales amount (projected from the linear regression model) with probability that a customer will buy the catalog ("Score_Yes"). Then the average gross margin was calculated by multiplying the expected revenue by 50%. Next, I subtracted the costs of printing and distributing of $6.5 from the average gross margin calculated in previous step to find the expected profit for each customer. At last, I added up all the expected profit for 250 customers in the list to determine whether the company should send the catalogs to customers.



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

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

According to the calculation, the expected profit from sending the new catalogs to these 250 customers is about $21,987.44 which exceeds our minimum acceptable expected profit contribution of $10,000.